THE UNIVERSITY OF
WARWICK

University of Warwick institutional repository: http://go.warwick.ac.uk/wrap

A Thesis Submitted for the Degree of PhD at the University of Warwick

http://go.warwick.ac.uk/wrap/3832

This thesis is made available online and is protected by original copyright.

Please scroll down to view the document itself.

Please refer to the repository record for this item for information to help you to cite it. Our policy information is available from the repository home page.

Study of the complete genome sequence of *Streptomyces scabies* (or *scabiei*) 87.22

Alice Morningstar Yaxley

Thesis submitted in partial fulfilment of the requirements

for the degree of

Doctor of Philosophy

University of Warwick,

Department of Biological Sciences

Sixth Month 2009

Table of Contents

List of Tables and Figures

Acknowledgements

Genome sequencing as described in **Methods 2.2** was carried out by *Streptomyces scabies* sequencing group, Wellcome Trust Sanger Institute (WTSI) in collaboration with United States Department of Agriculture via grant to Prof R. Loria. A**utomated annotation transfer pipeline** described in **Methods 2. 3** initiated on the completed sequence by Dr S. D. Bentley. **Confirmation of products** encoded in the genome was undertaken by collaborators Dr L. Song and Prof. G. L. Challis and Prof R Loria and Dr. R. F. Seipke as stated in the text.

Many thanks to Dr S. D. Bentley for his very generous contributions of tutorial time in genome annotation and interpretation. I was very lucky to have the opportunity to work at Wellcome Trust Sanger Institute to annotate the genome and to learn genome interpretation skills. Many thanks to Dr S. D. Bentley for hosting me at WTSI and also to Prof. J Parkhill, Dr L. C. Crossman, Dr M. Sebaihia, Dr N. R. Thomson, Dr M. T. Holden, and Dr A. M. Cerdeño-Tárraga, Dr T. Carver, N. Peters, and G. Vernikos of the Pathogen Sequencing Unit, WTSI for computational resources, advice and programming support.

Thanks to my supervisors, Prof. E. M. H. Wellington, Department of Biological Sciences and Dr R. G. Allaby, Warwick HRI, both at University of Warwick.

Thanks to Prof. G. L. Challis for explaining the structure-based predictive model and providing many hints on the interpretation of sequence data implicated in biosynthesis of complex natural products. Many thanks also to those others who have advised me with project and data management, informatics tools, programming, and chemical biology including L. P. Crisp, Dr A. Marsh, Prof E. Nesbitt, D. Yaxley, S. R. Fraser, M. Waddilove, Dr B. Tiwari, Dr D. Field, A M Webber and E. C. Speirs.

Thanks to those past and present members of Prof Wellington's research group including Dr L. A. Calvo-Bado, Dr W. H. Gaze, Dr. M. Krsek, Dr S. Bryan, Dr A. Ul-Hassan, P. Laskaris-Bountourakis and S. Griffiths who have been good academic colleagues.

I could not have completed this project without the support of my husband D. Yaxley and many good friends and the co-operation of my daughter E. S. Yaxley. Very many thanks for the generous help of my parents and friends in looking after baby so I could work: Gillian and William Waddilove, Leif Burrough, and Natalia Graňa. My thanks also to the dentists who sorted out my sore tooth, as promised.

Declaration

This thesis is the candidate's own work

It contains work based on collaborative research and work published in other documents as follows. No material from this thesis has been previously submitted for a degree examination at this or any other University.

Publications

Methods developed in this work were used by the author to study a novel non-ribosomal peptide synthetase (NRPS) system in the genome of *Pseudomonas fluorescens* SBW25 in the following work recently published (pdf attached, Appendix 2):

Silby, M., A. Cerdeno-Tarraga, G. Vernikos, S. Giddens, R. Jackson, G. Preston, X.-X. Zhang, C. Moon, S. Gehrig, S. Godfrey, C. Knight, J. Malone, Z. Robinson, A. Spiers, S. Harris, G. Challis, A. M. Yaxley, D. Harris, K. Seeger, L. Murphy, S. Rutter, R. Squares, M. Quail, E. Saunders, K. Mavromatis, T. Brettin, S. Bentley, J. Hothersall, E. Stephens, C. Thomas, J. Parkhill, S. Levy, P. Rainey and N. Thomson (2009). Genomic and genetic analyses of diversity and plant interactions of Pseudomonas fluorescens. *Genome Biol* **10**(5): R51.

The author collaborated on a chapter for the 2006 symposium volume of the Society for General Microbiology which contributed to the author's work in development of methods for studying gene clusters, also during this project (pdf attached, Appendix 3):

Morningstar, A., W. H. Gaze, S. Tolba and E. M. H. Wellington (2006). Evolving gene clusters in soil bacteria. In *Prokaryotic Diversity, Mechanism and Significance*. N. A. Logan, H. M. Lappin-Scott and P. C. F. Oyston. Cambridge, Cambridge University Press**:** p201-222

(Declaration continues overleaf.)

Declaration continued – details of collaboration

Genome sequencing

The genome was sequenced and assembled and a standard pipeline for automated transfer of annotation in preparation for the author's curation work was operated at Pathogen Sequencing Unit, Wellcome Trust Sanger Institute as described in the text.

The author annotated the complete genome sequence of this organism, by inspecting, correcting, and adding data as described in the Methods chapters **2** and **3**.

This annotation of the genome has already been used in genome project collaborators' research in the following works:

Seipke, R. F. and R. Loria (2009). Hopanoids are not essential for growth of *Streptomyces scabies* 87.22. *J Bacteriol*.

Loria, R., D. R. Bignell, S. Moll, J. C. Huguet-Tapia, M. V. Joshi, E. G. Johnson, R. F. Seipke and D. M. Gibson (2008). Thaxtomin biosynthesis: the path to plant pathogenicity in the genus *Streptomyces*. *Antonie Van Leeuwenhoek* **94**(1): 3-10.

Seipke, R. F. and R. Loria (2008). *Streptomyces scabies* 87.22 possesses a functional tomatinase. *J Bacteriol* **190**(23): 7684-92.

Joshi, M. V., D. R. Bignell, E. G. Johnson, J. P. Sparks, D. M. Gibson and R. Loria (2007). The AraC/XylS regulator TxtR modulates thaxtomin biosynthesis and virulence in *Streptomyces scabies*. *Mol Microbiol* **66**(3): 633-42.

Functional Genomics

The author identified gene clusters, annotated those clusters by describing the coding sequences, the conserved domains identified within the coding sequences, similarities to previously known proteins, and made predictions of complex natural products possibly produced, based on the genome sequence.

Genome project collaborators Prof. G. L. Challis and Dr L. Song (Department of Chemistry, University of Warwick) undertook work to confirm several products by culturing the sequenced strain in iron-free medium and producing analytical data by liquid chromatography and tandem mass spectrometry (LC-MS/MS) and by High

resolution time of flight mass spectrometry (HR-TOF-MS) methods as indicated in the text and that work has not otherwise been published yet.

Genome project collaborators Prof R. Loria and Dr. R. F. Seipke investigated production of hopanoids following my annotation of the cluster and prediction that hopanoids could be produced and published the results of that work (Seipke and Loria 2009).

Summary

A study of the complete genome sequence of *Streptomyces scabies* 87.22, a common causative agent of scab disease of tubers including potato (*Solanum tuberosum*), is described. This work includes annotation of the genome and in-depth description of gene clusters likely to encode biosynthetic pathways for complex natural products and not also found in either "*Streptomyces coelicolor*" A3(2) or *Streptomyces avermitilis* MA-4680.

Twenty-eight gene clusters were identified as likely to encode enzymes for the biosynthesis of complex natural products. Substances predicted by this work, not previously known to be made by *S. scabies* 87.22, were confirmed by collaborators as products - desferrioxamines, germicidins, and hopene. Of the clusters identified, fourteen gene clusters are not conserved in the other two streptomycete genome sequences for which comparisons have been undertaken. The *Streptomyces* genus is a reservoir of producer organisms from which many complex natural products of therapeutic importance have been isolated. These findings suggest that the cargo of cryptic and silent gene clusters amongst other members of this genus may add significantly to previous estimates of undiscovered bioactive natural products.

Methods developed in this work could enable other researchers to rapidly identify gene clusters likely to encode enzymes involved in biosynthesis of complex natural products from complete genome sequences. De-replication is a problem for approaches to drug discovery based on activity screening and isolation of wild producer organisms. Computational methods in this work allow rapid de-replication of gene clusters following sequencing which may lead to discovery of many new natural products with therapeutic benefit.

Sequences predicted to be involved in scab disease pathogenicity are not found in only one 'pathogenicity island' location as expected, but at several loci. Two possible mechanisms were identified from sequence data which it is suggested could be involved in regulation of pathogenicity traits: an MbtH-like protein family and an iron box sequence likely to be triggered response to low iron conditions.

Abbreviations used in the text

AT acetyltransferase, domain involved in polyketide biosynthesis.

bp number of nucleotide base pairs.

CDS coding sequence (ORF encoding a gene).

DH dehydrogenase, reductive loop domain involved in polyketide biosynthesis.

ER enoylreductase, reductive loop domain involved in polyketide biosynthesis.

INSDC.International Nucleotide Sequence Database Consorrtium, the agreement between the international hosts of nucleotide sequence data; DBJ, GenBank, EMBL.

kbp thousand base pairs.

KR ketoreductase, reductive loop domain involved in polyketide biosynthesis.

KS ketosynthase condensation domain involved in polyketide biosynthesis.

nr non-redundant – a subset of sequence deposited in INSDC.

NRPS non-ribosomal peptide synthetase, one of the main kinds of biosynthetic multienzyme systems.

PAI pathogenicity island.

PKS polyketide synthase, one of the main kinds of biosynthetic multienzymes systems.

1 Introduction

This work is a study of the complete genome sequence of the plant pathogen *Streptomyces scabies* 87.22. *S. scabies*, also known as *S. scabiei*, is one of a number of scab-causing organisms that cause mainly superficial damage to potato (*Solanum tuberosum*) and sweet potato (*Ipomoea batata*) tubers, and other root crops. *Streptomyces scabies* infection does not produce major crop losses in the U.K., but in much more arid regions the pathogenic effect is more severe (Loria *et al.* 1997; Wilson *et al.* 1999; Hill and Lazarovits 2005). Economic losses happen because the appearance of the tubers is affected by scab infection which decreases the marketability of tubers and increases the mass of tuber lost in peeling during industrial processing (Loria 1991).

Complete genome sequences have been determined and are freely available for two other organisms in the genus, "*Streptomyces coelicolor*" A3(2) (Bentley *et al.* 2002) and *Streptomyces avermitilis* MA-4680 (Ikeda *et al.* 2003). Both of these are non-pathogenic, and have been used extensively in this work for comparison with *S. scabies* 87.22. The complete genome sequence of *Streptomyces griseus* subsp *griseus* NBRC 13350 is now available (Ohnishi *et al.* 2008) but has not been used extensively for comparison because this work was largely undertaken prior to public release of the *S. griseus* complete sequence.

1.1 Streptomyces *genomes*

1.1.1 A genus of producer organisms

Streptomyces scabies 87.22 belongs to a genus known for antibiotic production. This genus is the origin of producer organisms accounting for perhaps 70% of the antibiotic compounds so far discovered (Berdy 2005). Streptomycetes are found ubiquitously in soil (Wellington and Toth 1994) and most humans have encountered them even without knowing because streptomycetes are thought to be the major producer of the volatile chemical odour characteristic of earth, geosmin (Bear and

Thomas 1964). Most streptomycetes are thought to live as saprophytes in soil, secreting sets of extracellular enzymes to break down recalcitrant polymers from plant material (Hodgson 2000), especially lignin and cellulose (Schlatter *et al.* 2009).

Organisms belonging to the *Streptomyces* Waksman and Henrici 1943 (Approved Lists 1980) genus are Gram positive and have a complex lifecycle with both spores and colonies of branching vegetative mycelium. Hydrophobic exospores are a dispersal form, germinating in appropriate conditions to form hyphae which branch and divide to form substrate mycelium. Hyphae elongate without complete division of cells to form a mycelium, laying down nucleoid bundles at intervals. Morphological differentiation begins with rearrangement of the cytoskeleton, cell wall assembly, cell division and chromosome segregation mechanism of subsequent cells (Flardh and Buttner 2009). Aerial hyphae grow and segment into compartments to form spore chains from which spores are released as they ripen (Glauert and Hopwood 1961).

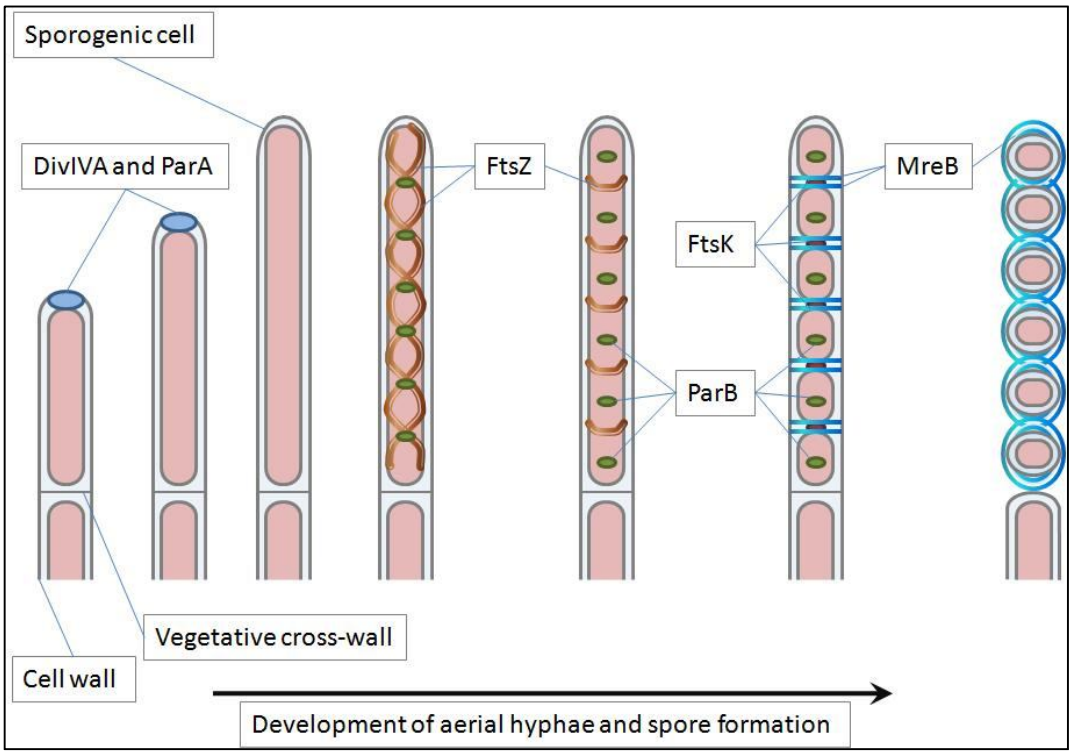


Figure 1-1 Morphological development in streptomycetes. Cell-wall assembly and cell division: Aerial hyphae grow by tip extension; arrest of growth; FtsZ helical filaments are remodelled into Z rings directing septation; ParB assembles at chromosomal OriC regions. Septal ingrowth starts over unsegregated nuclear material; FtsK DNA translocase, targeted to division sites, helps to clear DNA from closing septa. Prespores assemble thick spore walls; MreB localizes to septa and then surrounds the developing spore; nucleoids condense in maturing spores. Figure adapted from Flardh and Buttner 2009.

Organisms of the *Streptomyces* genus are relatively slow-growing, but otherwise easy to isolate and culture in laboratory conditions. The fastest time for cell mass to double seems to be around 2 hours during exponential growth, and 2-7 days are required for complete sporulation (E.M.H.Wellington pers. comm.). These times are likely to be much slower in the natural environment, from necessary adaptions when temperature, nutrient supply and hydration fluctuate.

Unresolved taxonomy of Streptomyces *genus*

Taxonomy is important because it undergirds the selection of organism for research purposes. It is necessary to establish how organisms are related to each other in order to determine how widely the implications of any discovery apply. Horizontal gene transfer, gene duplication and deletion events and chromosomal rearrangements are all thought to have contributed to the evolutionary descent of bacteria into current genome structures (Ventura *et al.* 2007) and these can complicate the task of describing evolutionary descent relationships amongst the *Streptomyces* genus. The morphology of spore chains and texture of spore surfaces have been used to identify species amongst organisms of the *Streptomyces* genus (Pridham and Lyons 1965). Subspecies have been amongst those with the same spore chain and spore surface types using criteria such as carbon utilisation, production of pigments in vegetative hyphae or aerial mycelium, and production of or resistance to antibiotics (Pridham and Lyons 1965).

At least 552 bacterial species have been validly described in the *Streptomyces* genus according to taxonomists' rules (Euzeby 1997; Euzeby 2009). Subgroups have been described within the genus (Williams *et al.* 1983; Anderson and Wellington 2001), but scab-causing strains have been discovered in several (Bramwell *et al.* 1998; Bukhalid *et al.* 1998). *S. scabies* was assigned to cluster 3, atroolivaceus group, in numerical taxonomy of the *Streptomyces* genus (Williams *et al.* 1983), but appears to be missing from the description of groups in Bergey's manual of systematic bacteriology (Williams *et al.* 1989). *S. scabies* has frequently been reported as belonging to the *diastatochromogenes* group (Locci 1994; Bukhalid *et al.* 1998).

Traits such as antibiotic and pigment production have been used to classify strains, but such traits may be subject to horizontal transfer (Alarcon-Chaidez *et al.* 1999;

Egan *et al.* 2001; Bukhalid *et al.* 2002). Hence an approach based on the molecular evolution of a highly conserved trait is necessary. Phylogenetic trees reconstructed from the DNA sequence encoding 16S ribosomal RNA (Woese *et al.* 1990) are useful but there are too few base changes to resolve the necessary number of branches to distinguish the large variety of organisms within the *Streptomyces* genus (Gao *et al.* 2006).

Some researchers have used "operational taxonomic unit" (OTU), groupings of 16S sequences greater than 99% identical, to cluster streptomycetes retrieved from soil samples (Davelos *et al.* 2004); others have suggested use of a 120 base pair subsection of 16s rDNA to cluster streptomycetes (Kataoka *et al.* 1997). Several approaches to phylogenetic classification depend on obtaining sequence data for other genes. A set of 233 genes have been identified as unique and apparently conserved in function amongst Actinobacteria (Gao *et al.* 2006), the phylum in which the *Streptomyces* genus is found. Presumably an approach selecting from amongst these might be illuminating for resolving evolutionary relationships in the Phylum.

Phenotype and 16S ribotype are not reliably correlated (Baines *et al.* 2007) probably due to frequent horizontal transfer events (Egan *et al.* 1998; Wiener *et al.* 1998; Tolba *et al.* 2002). Reconstruction of evolutionary history through phylogeny relies on vertical descent so a non-transferred core genome, which could be informative for taxonomy, must be reliably distinguished from the potentially transferrable accessory genome. This core genome can be identified through a combination of sequence-based methods (Philippe and Douady 2003) and is confirmed through comparisons of expression patterns (Callister *et al.* 2008).

Thus a basis exists for taxonomic approaches to resolve relationships with the Streptomyces genus, but much work remains to be undertaken. The evolutionary structure of the genus is an important question as it might enable researchers to explore the extent to which the diversity of the genus has already been sampled in the complete genome sequences projects so far undertaken. However, such approaches rely on the availability of high quality complete sequences for genes of interest.

1.1.2 Recognizing *Streptomyces scabies* or *scabiei*

Strains classified by phenotype as *S. scabies* or *scabiei* grow brownish pigmented mycelium, produce gray spores from spiral spore chains, and are producers of melanoid pigments (Lambert and Loria 1989). The proper name of the species according to Names with Standing in Nomenclature (Euzeby 1997; Euzeby 2009) is *Streptomyces scabiei* corrig. (*ex* Thaxter 1892) Lambert and Loria 1989, sp. nov., nom. rev.. A request for an opinion from the International Committee on Systematics of Prokaryotes is pending (Lambert *et al.* 2007), on the grounds that the species epithet *scabies* was correctly used before *scabiei* was coined. Revisions have been proposed (Truper and DeClari 1997; Young and Euzeby 2008) to the International Code of Bacterial Nomenclature (Lapage *et al.* 1990) which may assist with this decision. The name *scabies* is the one generally known amongst plant pathologists and agriculturalists (D. Lambert pers. comm.) and has been used in this work in anticipation of the revision, a decision about which may be made in 2011.

Sequenced strain

The sequenced strain *Streptomyces scabies* 87.22 was isolated in 1987 by R. Loria from a deep-pitted scab lesion on a potato (cultivar Russet Burbank) from a commercial production field in the state of Wisconsin, USA. It has been stored as a spore suspension in 20 % v/v glycerol at -80 ^{o}C. 87.22 may be considered as the same species as the *S. scabies* type strain (R. Loria pers. comm.).

The type strain most closely related with ‘standing in nomenclature’ – formal taxonomic definition agreed by the International Committee for Systematic Bacteriology (ICSB) - is *Streptomyces scabiei* corrig. (*ex* Thaxter 1891) Lambert and Loria 1989 (Lambert and Loria 1989),

1.1.3 Previous *Streptomyces* genome projects

“Streptomyces coelicolor” A3(2)

The complete genome sequence of this organism was completed in 2002 (Bentley *et al.* 2002) and at the time was the largest completely sequenced bacterial genome with nearly 8.7 M base pairs. The genome was discovered to have 20 or more gene

clusters encoding enzymes for biosynthesis of complex natural products including undecylprodigiosin, actinorhodin, methylenomycin, and the calcium-dependent antibiotic. There appeared to be several distinct sets of genes likely to encode growth-stage dependent metabolic capacity, similar to tissue-specific capacity in eukaryotic genomes (Bentley *et al.* 2002).

The name of this organism is properly typed in double quotes because although it is almost ubiquitously known as "*Streptomyces coelicolor*", that name belongs to a different organism under taxonomic rules. "*Streptomyces coelicolor*" A3(2) belongs to the bacterial species *Streptomyces violaceoruber* (Waksman and Curtis 1916) Pridham 1970 by phenotype, as the aerial mycelium appears ash gray in colour, the spore chains have a spiral shape and spores appear smooth (Kutzner and Waksman 1959).

Streptomyces avermitilis MA-4680

The second published complete genome sequence in the genus was that of *Streptomyces avermitilis* MA-4680, a producer of avermectins which are important antiparasitic medicines. The complete genome sequence was determined to consist of over 9 M base pairs in a single linear chromosome, mean GC content 70.7% and annotation revealed 7,574 coding sequences. The core/arms structure was similar to that of "*S. coelicolor*", with 6.5 M base pairs in the core region showing all known essential genes and broadly conserved gene order, and arms in which no obvious synteny is seen. The capacity encoded in the genomes of both *S. avermitilis* MA-4680 and "*S. coelicolor*" A3(2) for biosynthesis of complex natural products is summarized in a recent review (Donadio *et al.* 2007)

Other Streptomyces genomes

Several other genomes of streptomycetes have been sequenced, but at the time the analysis began those sequences were not available for use. The complete genome sequence of *Streptomyces griseus* IFO 13350 (Ohnishi *et al.* 2008) was published after this work was at an advanced stage and hence is not included in comparisons. Several Streptomyces genomes have been sequenced by the Broad Institute and draft sequences have been released for these organisms: *S. roseosporus* NRRL 11379, *S. clavuligerus* ATCC 27064, *S. sp.* SPB78, *S. roseosporus* NRRL 15998, *S. sp.* C, *S.*

pristinaespiralis ATCC 25486, *S. sp.* E14, *S. lividans* TK24, *S. griseoflavus* Tu4000, *S. sviceus* ATCC 29083, *S. viridochromogenes* DSM 40736, *S. sp.* SPB74, *S. albus* J1074, *S. ghanaensis* ATCC 14672, *S. hygroscopicus* ATCC 53653, *S. sp.* Mg1, *S. sp.* AA4, *S. flavogriseus* ATCC 33331. [Source: Actinomycetales group **http://www.broadinstitute.org/annotation/genome/streptomyces_group/GenomeStats.html** accessed 2009-12-01.] Complete 16S sequences for these organisms has not been released and hence it is not yet possible to do more than guess at the coverage of the genus gene-space that has so far been achieved.

1.1.4 General features of *Streptomyces* genomes

Streptomycete genomes so far sequenced have several common features and common challenges for researchers. The genomes are large, almost twice as large as related organism such as those of the *Mycobacterium* genus (Bentley *et al.* 2002). There appear to be six copies of the ribosomal RNA operon in each *Streptomyces* genome so far sequenced and such large repeats are a challenge for correct assembly. These exceptionally large genomes may arise from a necessity for diverse genetic resources to survive in the extremely spatially and temporally variable conditions of the niche of streptomycetes that is probably most common, saprophytic life in soil. Genetics of the streptomycete linear chromosome and its replication have been recently reviewed (Hopwood 2006) so only a few key points are presented here.

Core/arms structure

Comparison with related genomes illuminated organization within the genome of "*S. coelicolor"* A3(2). Unconditionally essential genes appeared to be restricted to a 6.4 M base pair "core" region, with two "arm" regions either side of it containing predominantly genes for traits considered conditionally adaptive (Bentley *et al.* 2002). This structure was discovered to be conserved in the genome of *S. avermitilis* MA-4680 (Ikeda *et al.* 2003).

Conservons

Conservons were first identified in "*S. coelicolor"* A3(2) (Bentley *et al.* 2002) from the repeated discovery of a thirteen sets of four coding sequences with the same conserved domains. This group are thought to encode membrane-associated

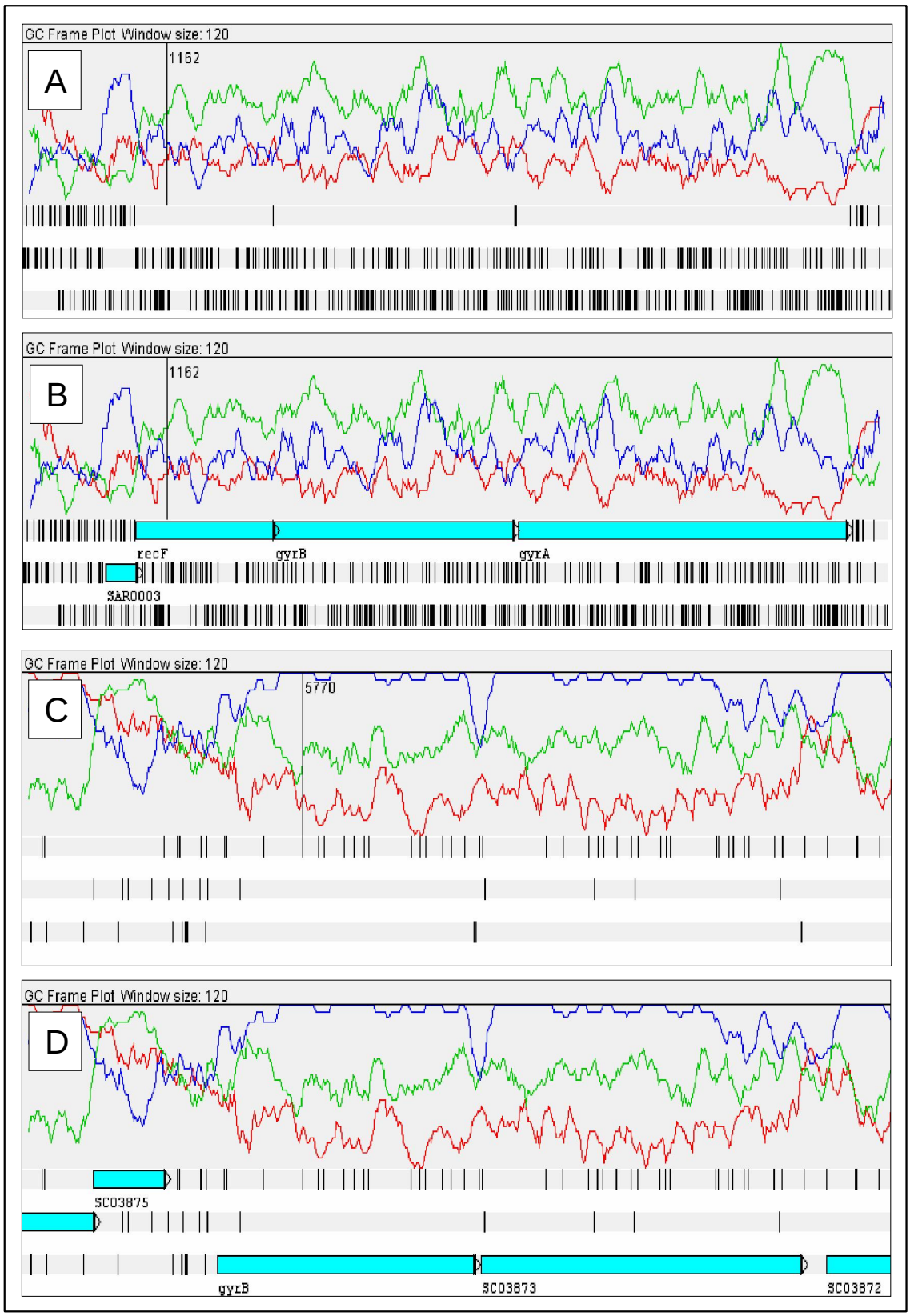


Figure 1-2 Illustration of composition-biased DNA and the coding sequence problem (A) *gyrB* and nearby coding sequences in *Staphylococcus aureus* MRSA252 (Holden *et al.* 2004), which has genomic G+C content 32% (B) shows positions of coding sequences, in the few open reading frames. (C) shows related sequence in *"Streptomyces coelicolor"* A3(2) (Bentley *et al.* 2002), G+C content 72%, showing greater number of open reading frames, (D) showing positions of coding sequences. Figures generated using Artemis (Rutherford *et al.* 2000).

signalling heterocomplexes with similarity to the eukaryotic G-coupled receptor

system (Komatsu *et al.* 2006). One of these conservons has a role in development of aerial mycelium in *"S. coelicolor"* and *S. griseus* (Komatsu *et al.* 2003) – others by similarity may play a role in sensing and responding in other conditions or stages of differentiation.

Linear genomes with protein-bound telomeres

All *Streptomyces* genomes so far sequenced are linear (Lin *et al.* 1993) with protein-bound telomeres (Yang *et al.* 2002; Bao and Cohen 2003). These protein-bound ends are covalently linked so the genome is thought to be effectively circular as a structure. This appears to be the usual form, although various circularized forms also arise (Redenbach *et al.* 1993; Leblond *et al.* 1996; Kameoka *et al.* 1999).

High G+C content

Genetic material from streptomycetes normally has a biased composition with more than 70% guanine/cytosine (G+C) nucleotide pairs. This composition makes the genetic material more difficult to handle, probably due to the order of magnitude greater thermodynamic stability of G+C base pairs versus adenine/thymine (A+T) base pairs (Pranata and Jorgensen 1991). The composition bias in theses genomes assists with identification of coding sequences (genes), because the third 'wobble' position of any amino acid codon in the genome has a greatly increased chance of being G or C, which gives characteristic FRAME plots (Bibb *et al.* 1984; Ishikawa and Hotta 1999).

The utility of FRAME analysis compensates for the additional difficulty of identifying the correct coding sequence from the many open reading frames found due to the bias in base composition. Because the STOP codons Amber (TGA) Ochre (TAA) and Opal (TAG) are A+T rich they are much less likely to be found in these highly biased genomes by chance alone than in the genomes of organisms without base composition bias.

There are thus a great number of open reading frames (ORFs) in a streptomycete genome, relatively few of which are expected to encode functional genetic material. Hence the more accurate term "coding sequences" in used in this work as in others to describe genetic loci predicted to encode proteins or stable RNA, where in genomes

with unbiased base composition ORFs would be correct. **Error! Reference source not found.** illustrates the great difference this makes to the appearance of genomic DNA visualized in Artemis: positions where presumably homologous *gyrB* and nearby genes are encoded have a very different appearance in high and low G+C organisms.

Bacterial gene finding

A great density of coding sequences is observed in bacterial genomes. It has been suggested that deletion bias works against the maintenance of coding sequences without selective advantage in bacterial genomes (Mira *et al.* 2001). Whatever the explanation, one feature of this density is an arrangement of slightly overlapping coding sequences in streptomycete genomes, which are probably cotranscribed as operons though not necessarily cotranslated (Flint *et al.* 2002). This gene density (90% coding) makes the task of correctly identifying coding sequences rather different from gene finding in eukaryotic genomes, which may have less than 10% (Salzberg and Delcher 2004).

1.1.5 Features associated with horizontal transfer

Many examples exist in streptomycetes where horizontal transfer seems to have occurred (Egan *et al.* 1998; Wiener *et al.* 1998; Egan *et al.* 2001; Tolba *et al.* 2002) especially during evolution of sequences related to complex product biosynthesis. The mechanisms of such transfer are not fully elucidated, though several have been extensively investigated.

Mobile elements provide bacteria with the ability to acquire genetic material by horizontal transfer. At least some pathogenicity traits in scab disease are found to be mobilised through horizontal gene transfer (Bukhalid *et al.* 2002; Kers *et al.* 2005). Horizontal gene transfer can occur through a number of mechanisms in bacteria, typically enumerated as transformation – direct uptake of genetic material; conjugation – transfer between bacteria by direct cell to cell contact; and transduction - phage-borne gene transfer.

The classification of a particular element found in genomic material can be difficult since groups of researchers appear to use different terms and classifications

depending on their focus. It has been recommended that mobile elements should be classified by the presence of the rather simple modular elements found in them (Toussaint and Merlin 2002). Those authors suggest the use of three module categories: features associated with transfer of genetic material between locations within a cell, features associated with transfer of genetic material outside the cell, and stability features.

Plasmids are natural vectors often found as independently-replicating covalently closed circular forms. In various organisms plasmids have been found to carry genes with clear selective advantages to the host such as symbiosis genes, antibiotic resistance or xenobiotic breakdown, and such selective advantage is thought to be essential in maintaining plasmids for example in *Escherichia coli* populations (Gordon 1992). Some plasmids can integrate into host chromosomes and these can be identified by the presence of characteristic plasmid reproduction genes. Transfer functions involve at minimum *kil* and *kor* genes, *kor* meaning 'kill override' (Kendall and Cohen 1987), because it encodes the repressor of the *kil* transfer gene. Without the repressor, the transfer gene overexpresses and causes cell death. The names given to these genes in the different mobile elements seem to vary wildly so it is easier to recognize them by conserved domains (Figure 1-3) than by similarities to named genes.

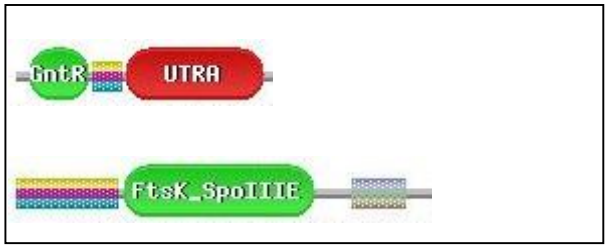


Figure 1-3 Bead-on-string view of conserved domain architecture of Pfam (Coggill *et al.* 2008) domains for plasmid *kor* and *kil* genes. Architecture of *korSA* (top), a kor gene from *S. amobofaciens* pSAM2 mobile element, and (bottom) *traSA*, kil gene from the same element.

The conserved domain called GntR after the first discovery of this kind of negative repressor (Bachi and Kornberg 1975) is described with Pfam model PF00392 which covers the N-terminal HTH region of GntR-like bacterial transcription factors. Kor appears to be a typical GntR-family regulator, as it has a ligand-binding domain at the C terminal of the primary sequence of the protein. UTRA is PF07702 and represents 'UbiC transcription regulator-associated' domain, and is a conserved ligand-binding domain. FtsK-SpoIIIE is the model PF01580, and it is found in a

wide variety of proteins including FtsK cell division protein from *Escherichia coli* and stage III sporulation protein E SpoIIIE in *Bacillus subtilis*, which is implicated in intracellular DNA transfer. Other distinctive sets of genes identify particular kinds of Streptomyces plasmid: pSAM2-type integrative elements have four 'spread' genes *spdABCD* genes as well as *xis* excisionase and *int* integrase genes (Possoz *et al.* 2001).

Bacteriophages are bacteria-specific viruses and can transmit genes between strains of bacteria by mistakes in repackaging (transduction). Some can integrate into host chromosomes at characteristic sites and the integrated forms are known as prophages, and can be recognized by the presence of conserved domains such as phage baseplate assembly (PF04717), coat protein (PF05357), tail fibre protein (PF04630), terminase involved in DNA packaging (PF05876) and so on as well as the integrase genes for example *rve* (PF00665). After integration prophages may become defective by losing essential viral reproduction genes and may gradually lose the identifiable genes for phage function.

Site-specific integration of bacteriophage λ in *Escherichia coli* occurs at a site named *attB*, between the *gal* and *bio* operons. This site is 30bp in length and contains a central region of 15bp where the recombination will take place (Gottesman and Weisberg 1971). Recombination is facilitated by both host and phage-encoded factors and the *attB* site is fused with a phage-encoded *attP* site to leave hybrid *attL* and *attR* sites surrounding the prophage (Abremski and Gottesman 1982). In "*Streptomyces coelicolor"* A3(2), sites similar to *attB* have been found as part of the integration site of prophage C31 and integrating plasmids (Combes *et al.* 2002).

Insertion sequence (IS) elements are the smallest autonomously replicating elements, and usually consist of short sections of DNA flanked by direct repeats and encoding just one protein, the transposase that catalyzes their mobility. They duplicate within genomes and can gain mobility for horizontal transfer by integration into conjugative elements. Insertion sequences can catalyse mobilization of host chromosomes, for example IS21 in *Escherichia coli* (Willetts *et al.* 1981), and after internal duplication in a genome IS elements may facilitate larger rearrangements by providing regions of identical sequence at which homologous recombination can begin. Several kinds of IS are found as part of larger elements, such as conjugative transposons.

Conjugative transposons have a recognisable structure with transposition elements flanking a cargo region, in which coding sequences can be transferred between strains. They were first discovered in Gram positive organisms *Enterococcus faecalis* and *Streptococcus pneumoniae* (Salyers *et al.* 1995). The majority of conjugative transposons so far studied use a tyrosine-recombinase and integrate in a non-site specific manner, with a preference for AT rich sequences (Osborn and Boltner 2002).

'Genomic island' seems to be used as a general term for regions with some indicators of mobility and includes toxin production islands such as in *Corynebacterium diphtheriae* (Cerdeno-Tarraga *et al.* 2003) as well as the one identified in *S. turgidiscabies* Car8 (Kers *et al.* 2005). Integrated mobile elements such as plasmids and conjugative transposons may become a favourable target for integration of further elements (Dobrindt *et al.* 2004) to form genomic islands. A newly integrated mobile element may have sections of DNA with no selective advantage to the host organism. Whereas natural selection will eliminate genomes in which integration has interrupted essential gene functions, integration of further mobile elements within the bounds of the acquired material is less likely to be selected against. Acquisition of insertion sequences by integrated mobile elements may lead to new mobility traits for the element such as additional recombination enzymes which could allow the element to reach new gene targets (Toussaint and Merlin 2002).

1.2 Biosynthesis of complex natural products

Although agriculturalists know *S. scabies* as the causative agent of scab, the *Streptomyces* genus is best known amongst microbiologists for the producer organisms from which antibiotics are harvested. Scab pathology is known to depend at least partly on production of thaxtomins (Loria *et al.* 2008), from a non-ribosomal peptide synthetase (NRPS) system. NRPS systems are better known as the biosynthetic origin of antibiotics such as penicillin (Schofield *et al.* 1997), and such antibiotics have recently been reviewed (Nolan and Walsh 2009).

Antibiotics are usually defined as those bioactive natural products having highly specific activity in low concentrations, with differential toxicity on host and pathogen – it is necessary that they kill the pathogen without killing the host. Although *S. scabies* is a pathogen, and not primarily an antibiotic producer, the importance of the genus as a reservoir of producer organisms makes the interpretation of this genome sequence a potentially important task for understanding the genetic context in which antibiotic production arise.

'Genome mining' for new natural products is an area of great research interest (Lautru *et al.* 2005; Corre and Challis 2007; Challis 2008a; Zerikly and Challis 2009), harvesting the results of the increasing volume of sequence data including discovery of new products from previously cryptic and silent gene clusters (Challis 2008b; Zerikly and Challis 2009). Comparative analysis of this genome could illuminate the context and frequency of gene clusters encoding the capacity for biosynthesis of these bioactive compounds.

Regulation of antibiotic production is obviously of great interest to researchers of producer organisms, and appears to be regulated in a sensitive fashion by many different mechanisms under active investigation by researchers around the globe (Bishop *et al.* 2004; Hara *et al.* 2009; Pan *et al.* 2009). There are several mysterious features, such as the activation of previously silent complex product biosynthesis capacity by supplying cloned fragments of DNA from other organisms (Jones, G. H. and Hopwood 1984; Fawaz and Jones 1994) Discoveries about regulation of pathogenicity traits in *S. scabies* could potentially throw light on related mechanisms in non-pathogenic streptomycetes, just as the results of investigations of regulatory mechanisms in other streptomycetes inform investigations into the mechanisms involved in pathogenicity regulation in *S. scabies*.

1.2.1 "Complex natural products"

Useful terminology for the bioactive natural products characteristic of this genus is not straightforward. Several ways of referring to them exist, none of which is entirely satisfactory. In this work 'complex natural products' has been chosen to refer to any potentially bioactive molecules including those previously referred to as secondary metabolites.

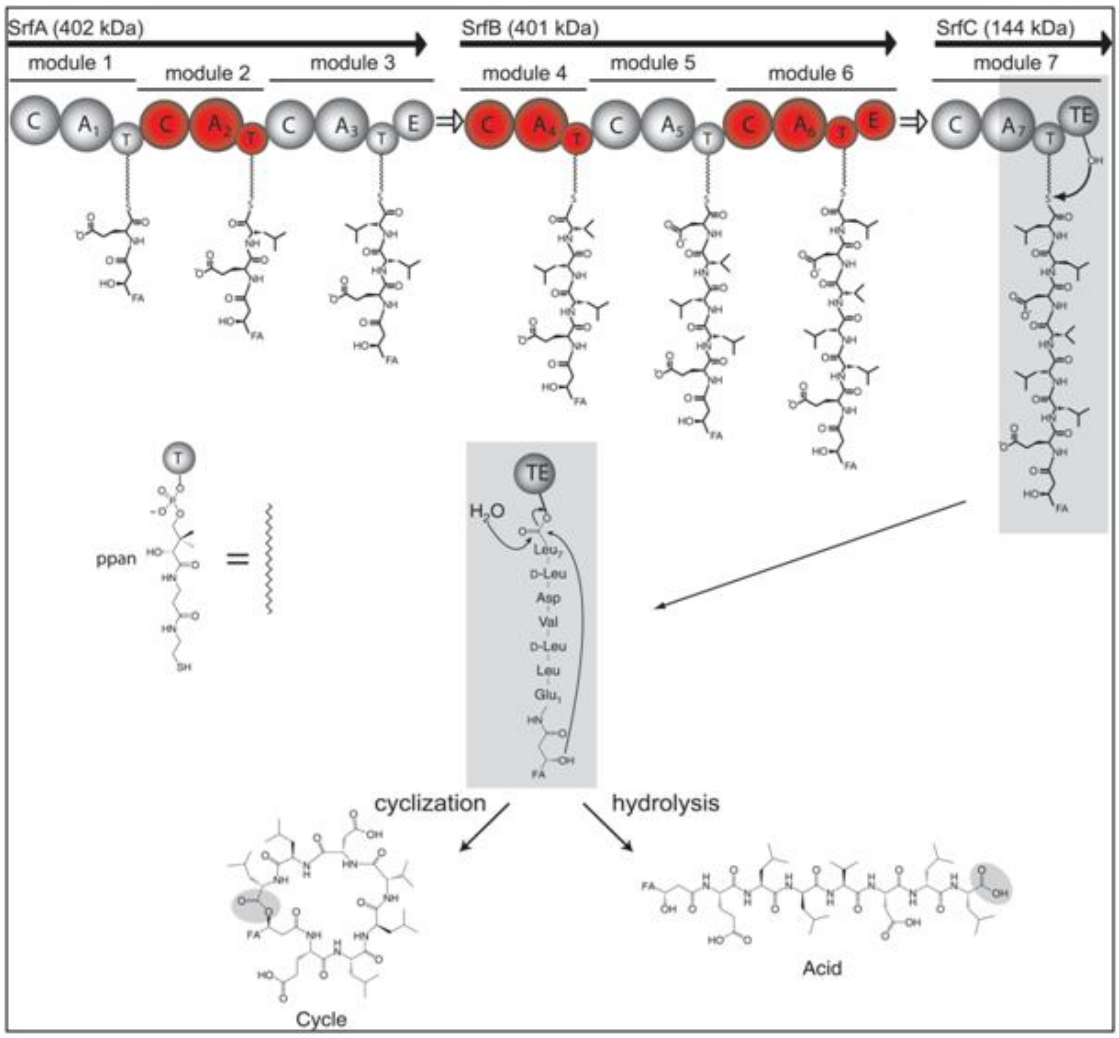


Figure 1-4 Illustration of a non-ribosomal peptide synthetase system. Surfactin synthetase from *Bacillus licheniformis*, reproduced from Sieber et al. 2003.

Some of these substances are likely to be adaptive compounds allowing growth in less common conditions. They include antibiotics, phytotoxins, and substances to discourage various kinds of predation; pigments; specialized ionophores for nutrient acquisition and redox homeostasis; molecules possibly involved in sensing a responding to the environment (Price-Whelan *et al.* 2006); and other compounds with biological activity likely to be essential in the organism's habitat though sometimes dispensable in laboratory culture.

Bioactivity is an important characteristic and could be indicated based on the maintenance of the gene clusters against the background of deletional bias (Mira *et al.* 2001). It should be borne in mind that the activity of many candidate molecules is unproven. The ecological role of these products seems to define the niches the organisms exist in, hence 'individualites' has been suggested (Piepersberg 2002),

and a similar term “idiolite” has seen some use in the past (Walker 1979; Walker 1988), but it seems better to avoid coining terms if a sufficient description can be found.

The capacity to produce these special compounds could also be defined by their inclusion in genomic content that varies between species (perhaps even responsible for speciation?) This flexible genomic content is frequently referred to as the “accessory” genome (Dobrindt *et al.* 2003; Sim *et al.* 2008), contrasted with the core genome which is conserved across a less closely related group of organisms, for example a higher taxonomic level. This qualification is an interesting one but threatens to open up the taxing questions of bacterial classification which must be left to more qualified researchers.

Such substances are usually classified with a molecular mass qualification, having low enough molecular mass to distinguish them from peptides encoded in DNA and produced by ribosomes, <3 kD (Berdy 2005). It might be sufficient to specify that they are the product of enzymes other than ribosomes, and not produced using an RNA template but examples exist of ribosomally manufactured short peptides modified by other enzymes to produce bioactive products (Hille *et al.* 2001; McClerren *et al.* 2006), so this also is also not an absolutely useful category.

The most common terminology for these compounds historically, “secondary metabolites”, appears to have arisen from the observation that complex natural products were often secreted *during the secondary phase of growth* alongside differentiation and sporulation processes, and *seemed dispensable* under laboratory conditions unlike those metabolic traits labelled “primary”. Production is often highly sensitive to the composition of media and not related to the doubling time of cells (Vining 1990).

However, several bioactive compounds are not linked to the stationary growth phase. The spore germination inhibitor germicidin is secreted as spores begin to germinate (Petersen *et al.* 1993; Song *et al.* 2006), and the odiferous compound geosmin is produced as organisms die (Zaitlin and Watson 2006). Furthermore, since capacity to produce several compounds which might otherwise be included, such as the iron-scavenging molecules known as desferrioxamines, do not seem to be dispensable

even in laboratory conditions as they must be exogenously supplied for growth (Yamanaka *et al.* 2005), neither qualification for the "secondary metabolite" terminology is clearly useful.

Complex has been used to refer to the complexity of their chemical structure, to distinguish them from simple sugars or lipids. This is also not perfect terminology as universally conserved bioactive molecules, ATP or coenzyme A might be classed as chemically complex. For the purpose of this work complex natural products are defined as potentially bioactive, under 3kD in mass, produced via enzymes other than ribosomes and not encoded in the core genetic material apparently conserved across the taxonomic superkingdom Bacteria.

1.2.2 Biosynthetic gene clusters vary

Coding sequences for biosynthesis of a particular function are frequently found in clusters on the chromosome. Such gene clusters could be defined by "co-ordinated regulation of a number of adjacent transcription units which may be found in both senses, strands, and any frame" (Morningstar *et al.* 2006). Clusters encoding enzymes for biosynthesis of complex natural products might be identified by conserved domains of enzymes with a role in biosynthesis of complex natural products such as non-ribosomal peptide synthetase (NRPS) systems and polyketide synthase (PKS) systems. Other systems such as NIS (Challis 2005) and other characterised biosynthetic pathways for complex natural products (Distler *et al.* 1992; Bursy *et al.* 2008) are more diverse and a combination of similarity-based and conserved-domain modelling approaches are required.

The search for genomic locations of these enzyme systems for complex natural products is complicated by the discovery that not all such systems are spatially clustered on genomes. Coding sequences involved in biosynthesis of massetolide A are found in two distinct clusters of genes (de Bruijn *et al.* 2008); viscosin genes in *Pseudomonas fluorescens* strain have also been found to have separated chromosomal locations (Braun *et al.* 2001).

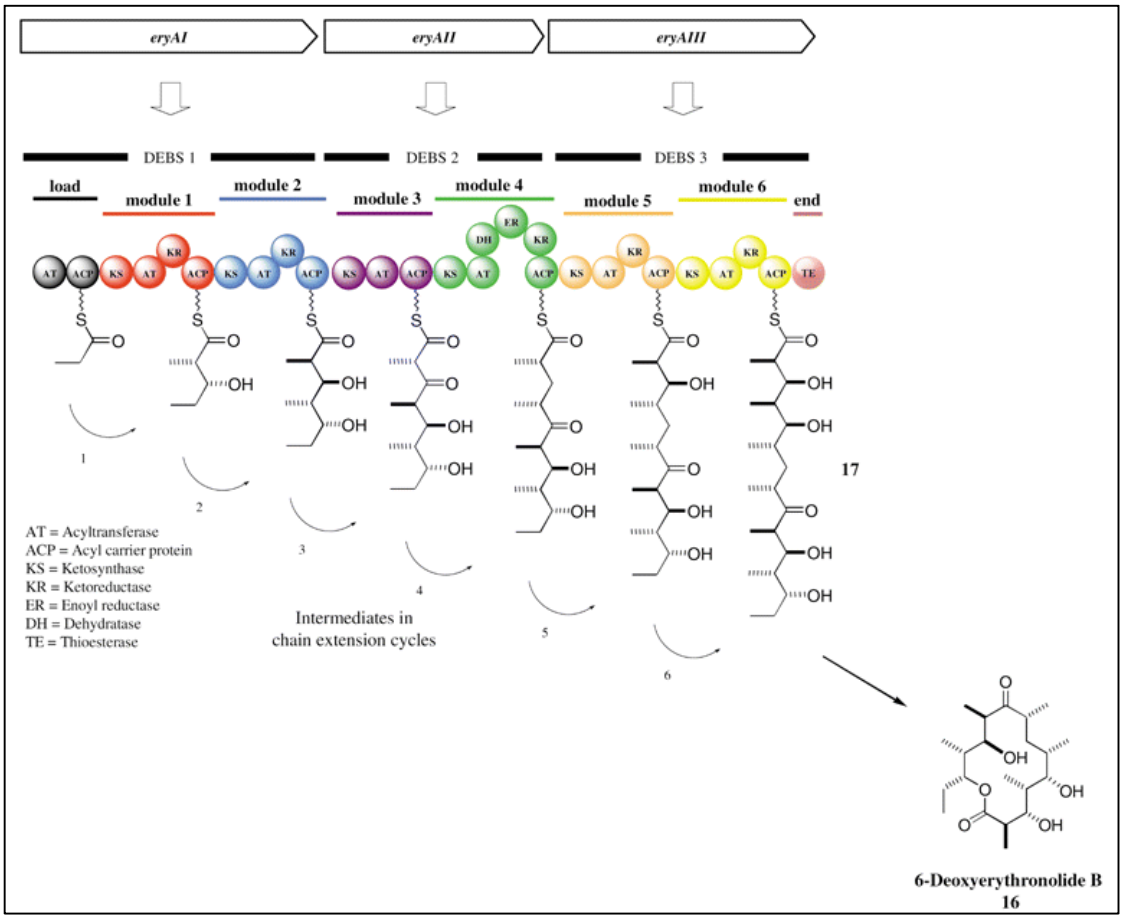


Figure 1-5 Illustration of polyketide synthase (PKS) system. Erythromycin synthase from Saccharopolyspora erythraea, illustration reproduced from Staunton *et al.* 2001

Problem of dereplication

Dereplication is a problem in the study of complex natural products: identical compounds have been rediscovered on multiple occasions by screening approaches (Berdy 2005). Sequencing offers an initial approach to the dereplication problem by confirming whether gene clusters encoding the biosynthetic enzymes are identical. However, this requires the highest level of attention to detail in sequence analysis. Just how similar do the sequences need to be to confidently predict that identical or different compounds are produced by the encoded enzymes? Active site residues need to be conserved in enzymatic domains to confidently predict function; and all domains or equivalent domains need to be present in the cluster.

1.2.3 Polyketide synthases

Polyketides synthase (PKS) systems (Figure 1-5) construct large molecules from carboxylic acid starter units according to a mechanism very similar to that of fatty

acid synthase (FAS) systems universally found in living cells. PKS systems generate a huge diversity of products by varying the degree of reduction applied to the extension units incorporated. One excellent review summarises the history of investigations so far into these systems (Staunton and Weissman 2001) and a forthcoming publication by is likely to provide an excellent introduction for a researcher new to natural product biosynthesis (Weissman 2009) judging by previous works by the same author.

1.2.4 Nonribosomal peptide synthetases

Nonribosomal peptide synthetases (NRPS) are usually large multienzyme proteins which catalyse biosynthesis of complex natural products including siderophores and antibiotics (Lautru and Challis 2004). NRPS enzymes join amino and other acids together (See Figure 1-4) based on a logic largely controlled by the structure of the multi-enzyme (Konz and Marahiel 1999). A relatively recent review can provide an introduction to the complications and the vast body of literature describing this important type of biosynthetic system (Finking and Marahiel 2004)

1.3 Streptomyces scabies *(or* scabiei*) the plant pathogen*

Scab-causing organisms appear to infect growing tubers through lenticels. Lesions expand as the tubers grow, and sporulation occurs in the lesion releasing at least some spores into soil. *S. scabies* spores appear to survive in soil, and it is thought these organisms grow and reproduce as a saprophyte (Loria 1991).

1.3.1 Thaxtomins

Early work on the basis of pathogenicity focused on the discovery of the phytotoxin thaxtomins, which appear to assist entry to plants by inhibiting cellulose production in expanding tissue (Scheible *et al.* 2003) and may define pathogenicity amongst scab-causing organisms(Loria *et al.* 2008). Thaxtomins directly affect plant cell wall deposition (Fry and Loria 2002) and cause hypertrophy in seedlings (Leiner *et al.* 1996). The cellulose biosynthesis inhibition effect of thaxtomins would appear to explain the developmental constraint of scab infection to expanding plant tissues (Loria *et al.* 2008).

Thaxtomins are dipeptide phytotoxins produced by *Streptomyces scabies* and related scab-causing species, *S. turgidiscabies*, *S. acidiscabies* (Wach *et al.* 2007) and *S. ipomoea* (King *et al.* 1994). Thaxtomin A is an important factor in virulence (Healy *et al.* 2000; Loria *et al.* 2008) and application of thaxtomins to immature potato tubers mimic scab symptoms (Lawrence, C. H. *et al.* 1990). Thaxtomin A production is not the only factor in scab pathogenicity however because not all mutants deficient in both melanin and thaxtomin A show reduced virulence as shown by common scab symptoms on potato tuber discs (Beausejour and Beaulieu 2004).

Control of thaxtomin biosynthesis has been investigated by a number of researchers. The regulatory protein TxtR, encoded in the biosynthesis gene cluster for thaxtomins, has recently been shown to bind cellobiose, the smallest oligomer of cellulose, and activate thaxtomin biosynthesis (Joshi, M. V. *et al.* 2007). Biosynthesis of thaxtomins may also be activated by a mixture of complex carbohydrates (Wach *et al.* 2007) and by fructose (el-Sayed el 2001). Thaxtomin biosynthesis appears to be inhibited by supplying the component amino acids tryptophan and phenylalanine (Lauzier *et al.* 2002). It also been suggested that plant cell lipids released by the activity of extracellular esterases triggers thaxtomin biosynthesis (Beausejour *et al.* 1999).

Thaxtomins are cyclic dipeptides which appear in the supernatant of flask cultures as a mixture of eleven variants differing by various hydroxyl and methyl groups (King *et al.* 2001). Thaxtomins contain an unusual nitro-indole moiety, likely to be derived from incorporation of gaseous nitric oxide (Kers *et al.* 2004). A chemical synthesis exists for thaxtomins A and B, (Gelin *et al.* 1993) also for thaxtomin C (King 1997).

Several streptomycetes other than scab-causing organisms may to be able to utilize thaxtomins as a carbon source (Doumbou et al. 1998). It has been suggested that this may account for the scab-suppression phenomenon. It could be argued however that the experimental methodology for utilisation of thaxtomins as sole carbon source does not take sufficient account of apparently autotrophic actinomycetes from environmental samples (Wellington and Toth 1994).

1.3.2 Phylogenetics of scab and suppressive organisms

Investigations into the organisms implicated in potato scab have been complicated by the apparent variety of causative organisms. Scabies and some other scab-causing appear to belong to the diastatochromogenes group within the *Streptomyces* genus (Takeuchi *et al.* 1996) by full-length sequences of 16S subunit rRNA (Edwards *et al.* 1989). Scab-suppressive streptomycetes also appear to belong to this subgroup (Eckwall and Schottel 1997)

Considerable phenotypic, genotypic, and pathogenic variation is found among streptomycetes involved in scab disease (Bramwell *et al.* 1998; Bouchek-Mechiche *et al.* 2000; Park *et al.* 2003) as well as amongst non-pathogenic streptomycetes sharing the lesion habitat (Doumbou *et al.* 2001). Sequences for the *rpoB* gene, encoding the DNA-directed RNA polymerase beta subunit [EC:2.7.7.6] equivalent to SCO4654 in *"S. coelicolor"* A3(2), as well as the rDNA sequences of the 16S ribosomal subunit, have been amplified in surveys of scab-causing organisms to help resolve the phylogenetic relationship (Mun *et al.* 2007; St-Onge *et al.* 2008). The evolution of pathogenicity in S. scabies has recently been reviewed (Loria *et al.* 2006).

Organisms of the *Streptomyces* genus are known to have a suppressive effect on pathogens in soil (Schlatter *et al.* 2009); this includes non-pathogenic *Streptomyces* apparently suppressing the growth of scab-causing strains(Eckwall and Schottel 1997). Scab-suppressing strains appear to be enhanced by green manure soil treatments (Wiggins and Kinkel 2005)

1.3.3 Extracellular esterase

An extracellular esterase was characterised and sequenced in *S. scabies* and has been suggested to be a virulence factor, possibly regulated by availability of zinc (Raymer *et al.* 1990). This enzyme could assist in pathogenicity by hydrolyzing bonds in suberin, a waxy coating on tubers, but it appears that involvement in pathogenicity has not been demonstrated. A similar esterase is found in the related non-pathogenic organism *S. diastatochromogenes* and the two enzymes appear to be similar in mechanism but differently regulated (Tesch *et al.* 1996). A promoter identified in front of the *S. scabies* coding sequence apparently has similarity to one of the

promoters of the agarase gene in *"S. coelicolor"* (Buttner *et al.* 1988; Tesch *et al.* 1996).

1.3.4 Nec1 necrosis factor

The *nec1* region has been identified as a factor in scab necrogenesis (Bukhalid and Loria 1997) and the presence of the 1.6kb DNA context of this gene has been used as diagnostic for the presence of pathogenic *Streptomyces* for agricultural use (Cullen and Lees 2007). Most *nec1*-containing strains tested in a trial of 43 pathogenic scab strains also produced thaxtomins (Bukhalid *et al.* 1998). The *nec1* sequence and the transposase-like sequence adjacent to it, *ORFtnp*, seem to transfer horizontally with high fidelity: strains with DNA-DNA hybridisation as low as 36% have been found to have identical *nec1* sequences (Bukhalid *et al.* 2002).

Transcriptional start sites of *nec1* appear to be affected by the presence of glucose and it is suggested that the Nec1 protein is a secreted virulence factor early in infection (Joshi, M. *et al.* 2007). Transfer of a 9.4kb fragment including the *nec1* region appears to be sufficient to allow necrotizing and colonisation of potato tuber slices (Bukhalid and Loria 1997). *"S. coelicolor"* transconjugants having acquired a very large region of sequence from the *S. turgidiscabies* PAI did not acquire pathogenicity traits, but may not have acquired this region (Kers *et al.* 2005). An insertion sequence IS1629 has been identified 3' of *nec1* and multiple copies of it have been found in *S. scabies* and related genomes (Healy *et al.* 1999).

1.3.5 Horizontal transfer of pathogenicity genes

Several scab-causing streptomycetes have been confirmed to have a highly conserved region around the *nec1* locus including a putative transposase (Bukhalid *et al.* 1998; Healy *et al.* 1999). Horizontal transfer of pathogenicity-related sequences has been demonstrated (Bukhalid *et al.* 2002). From partial sequencing of the *tomA* gene in scab causing isolated this is one of the highly conserved loci (St-Onge *et al.* 2008), hence is likely to be part of the transferred region.

1.3.6 A mobile pathogenicity island

A large pathogenicity island (PAI) has been identified (Bukhalid *et al.* 2002) in the related scab-causing organism *Streptomyces turgidiscabies* Car8 and complete sequencing of the whole genome of this related organism is underway (R. Loria pers. comm.). Some sequence data from the PAI, which may be as large as 660 000 base pairs in size, has been published (Kers *et al.* 2005). Data from published DNA from the pathogenicity island has been used in this work to identify conserved regions in the *S. scabies* 87.22 genome which may be involved in pathogenicity.

Since virulence traits in *S. scabies* 87.22 are known to be at least partially mobile (Kers *et al.* 2005), a survey of potentially mobile regions in the *S. scabies* 87.22 genome is included in this work. This survey has been compiled by identifying regions of difference with the other complete genome sequences available, the non-pathogenic streptomycetes, "*S. coelicolor*" A3(2) (Bentley *et al.* 2002) and *S. avermitilis* MA-4680 (Ikeda *et al.* 2003).

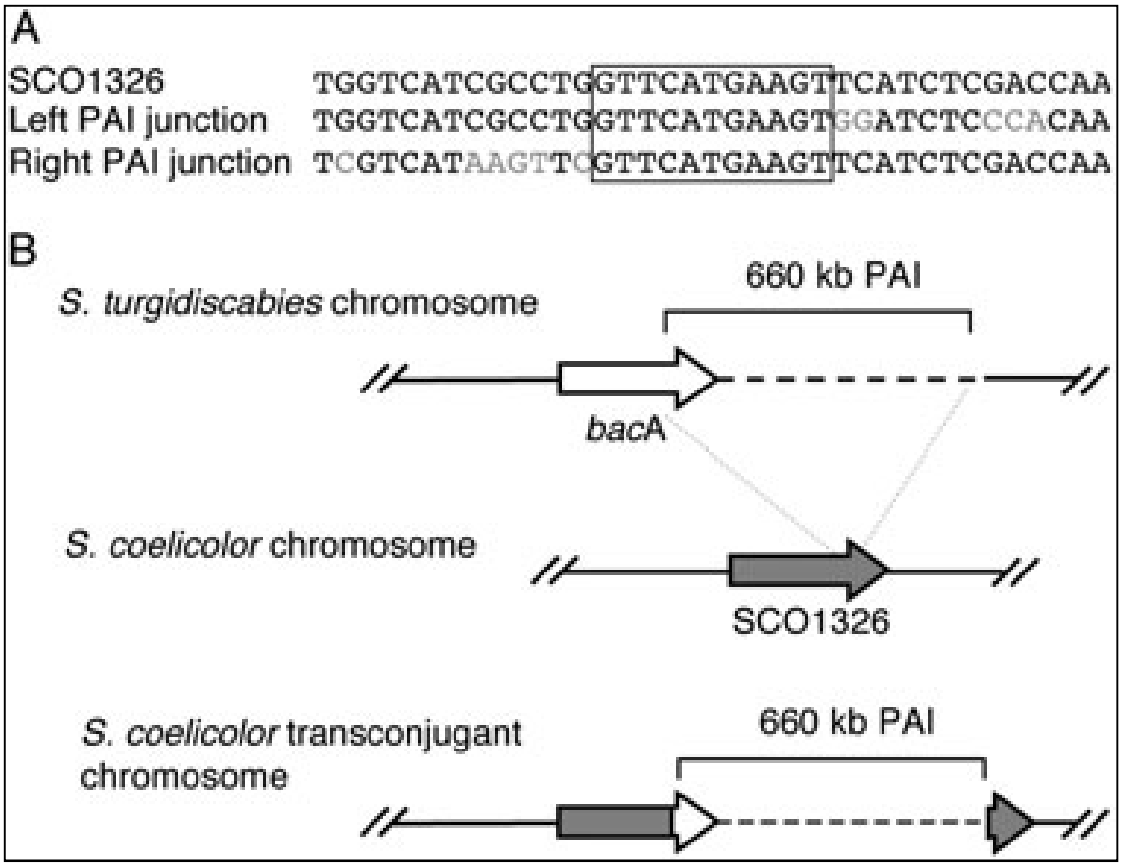


Figure 1-6 Integration site for pathogenicity islands in *"S. coelicolor"* transconjugant. Reproduced from Kers *et al*. 2005.

It has been determined that a large section of DNA can be transferred from scab-causing *S. turgidiscabies* Car8 to *S. diastatochromogenes* conferring the ability to necrotize potato tuber discs and produce thaxtomins (Kers *et al.* 2005). This transferable region has several features common to pathogenicity islands, including conservation of virulence factor sequence across organisms more distantly related by phylogenetic marked such as 16S ribomsomal DNA, and the presence of several mobility-related features such as multiple transposase-like and integrase-like genes (Kers *et al.* 2005) . Hybridization of probes for *nos* and *tomA* genes demonstrate that the transferred region from *S. turgidiscabies* Car8 to previously non-pathogenic strains contains the known pathogenicity loci. The nitric oxide synthase gene *nos* is associated with the gene cluster for biosynthesis of the phytotoxin thaxtomin, and *tomA* is located close to the *nec1* region(Kers *et al.* 2005).

The PAI appears to insert into a conserved motif in a coding sequence for an integral membrane protein of the bacitracin-resistance protein family (Kers *et al.* 2005). Sequencing of the proposed region of integration in transconjugants revealed the presence of a 10bp insertion site (Figure 1-6). Several copies of this 10bp sequence

are found in the "*S. coelicolor*" A3(2) genome but only the copy inside the bacitracin-resistance family protein is inserted into (Kers *et al.* 2005). The 3' terminus of the PAI contains approximately 126 base pairs flanking the truncated portion of SCO1326, and the 5' terminus of the island contains another member of the bacitracin resistance protein family.

None of the 15 "*S. coelicolor*" transconjugants found to have acquired PAI sequences acquired pathogenicity traits (Kers *et al.* 2005), so it has been suggested that the sequences contained on the PAI are necessary but not sufficient for pathogenicity (Loria *et al.* 2006).

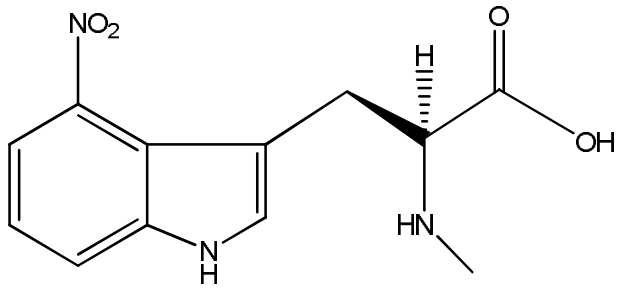
1.3.7 Saponinase?

A gene with homology to saponinases was discovered on the pathogenicity island of *S. turgidiscabies* (Kers *et al.* 2005), and it is functional in *S. scabies* 87.22 (Seipke and Loria 2008). Saponins are glycoalkaloids produced by plants which may act as a preformed defence against fungal pathogens (Jones, N. A. *et al.* 2005). Several fungal pathogens are known to utilize saponinases (Osbourn 1996). Saponinases are members of glycosyl hydrolase family 10 with the ability to deglycosylate saponin compounds to less toxic forms.

This gene may have a role in pathogenicity as degradation products of the saponin alpha-tomatine suppress host defence responses against the fungal tomato pathogen *Septoria lycopersici* (Bouarab *et al.* 2002). It is possible that the tomatinase in *S. scabies* 87.22 has been retained by natural selection for similar reasons. Other microbes also appear to utilize saponinases, notably the tomato pathogen *Clavibacter michiganensis* subsp. *michiganensis* NCPPB382 (Kaup *et al.* 2005).

1.3.8 Nitric oxide (NO)

A nitric oxide synthase is found in the gene cluster for thaxtomin biosynthesis, and *S. scabies* 87.22 has recently been shown to produce gaseous nitric oxide (Johnson *et al.* 2008). Although nitric oxide is a common signalling molecule in animal cells, and is part of defence signalling in plant cells, production in bacteria is uncommon (any others, which?).



1 R = CH_3 **2** R = CH_2CH_3

Nitric oxide inhibitors appear to reduce production of thaxtomin without affecting growth , and thaxtomin production can be partially restored after the nitric oxide synthase gene is knocked out by supplying gaseous NO (Wach *et al.* 2005). Two kinds of N-acylated 4-nitrotryptophans (**1** , **2** above) have been identified as non-phytotoxic exudates during *in vitro* production of thaxtomin A by *Streptomyces scabies (King and Lawrence 1995).* Chemical and biochemical studies have indicated that a 4-nitrotryptophan may be an intermediate in production of thaxtomins (King *et al.* 2003).

1.3.9 Concanamycins

Concanamycins are polyketide-sugar phytotoxins which affect the vacuolar H+ ATPase (Huss *et al.* 2002). Concanamycin A is used as a specific inhibitor of eukaryotic V-ATPases in pharmacological studies (Martinez-Munoz and Kane 2008). Concanamycins have toxic activity against fungi and yeasts (Kinashi *et al.* 1984), probably because they attack the V-ATPases. They may have therapeutic uses in humans (Dai *et al.* 2005). Although the toxicity of these compounds suggests they may affect the plant host the role of concanamycins in scab pathogenicity has not been demonstrated.

Concanamycins have an 18-membered macrolactone ring structure and an unusually folded sidechain which gives this class of compounds the name plecomacrolides (Insert structural synthetic information). Several organisms are known to produce concanamycins, including *S. scabies (Natsume et al. 2001), S. acidiscabies* (Natsume *et al.* 2001), *S. diastatochromogenes* (Kinashi *et al.* 1984) and *S. neyagawaensis.* The complete sequence of the cluster of genes responsible for biosynthesis of concanamycin A has been determined in *S. neyagawaensis (Haydock et al. 2005),*

which groups in the *diastatochromogenes* subgroup of the genus with *S. scabies* 87.22.

1.4 Annotation

Annotation refers to the process and product of interpretation of a genome. The simplest level of annotation is identifying the position of coding sequences. The nature of the coding sequences may be predictable by similarity-based comparison or from conserved domain architectures, but any level of annotation beyond the simplest is a guess (Parkhill 2002). Annotation is the assembly of a secondary data set based on inferences from the primary data set, the assembled DNA sequence. The task of annotationg involves assembling and integrating data from selected sources, assessing the importance and relevance of each for robust predictions about function and providing a preliminary interpretation of the meaning of the sequence data.

In several respects adequate annotation of microbial genomes remains a very difficult task. Bioinformatics skills are necessary to undertake the task and integrate the results drawn from the many databases which may be drawn on. In order to adequately identify and interpret traits encoded in the genome a good general knowledge of microbial physiology and biochemistry is required to annotate core functions. In addition, knowledge of the organism's particular biology is required to annotate species- and genus- specific traits correctly; and in the case of a pathogenic organism, it is necessary to know the specifics of previous research on the pathogen and the general state of research on host-pathogen interactions, to identify relevant features on the genome.

Annotation of a genome is the crucial step allowing interpretation of its contents and the route towards further investigations. A large volume of data can be generated but making sense of this data is just as reliant on well-designed, focused hypothesis-testing as non-genomic studies. Annotation has developed alongside sequencing technologies, but it may be that interpretation of genomic data is a bottleneck in the path of progress as sequencing technologies are increasingly scaled up (S. D. Bentley pers. comm.). Transferable advances in the interpretation stage of genomic science have the potential to unlock the vast and largely untapped diversity of environmental

organisms, but only if annotation is correctly applied and used with care. Challenges include picking out what is most relevant to the biology of the organism and exploring the extent to which deterministic results are possible from computational studies.

It is assumed that the annotation created as a result of this work will be treated as a first pass treatment rather than as authoritative. It is hoped that researchers using the sequence will use the lines of evidence provided to check how appropriate the annotation supplied is for their purposes. Researchers with specialist interest in any particular gene or cluster will rapidly surpass the level of detail possible for the annotator to apply to a complete genome sequence. It should be clear that annotation to create the genome file available for other researchers to work from is the beginning of the genome project's utility. Researchers have already begun to explore the implications of the annotation described in this work (Loria *et al.* 2008; Seipke and Loria 2008; Seipke and Loria 2009) and it is expected that more will be published shortly including the public release of the whole genome analysis.

Data curation

Similarity searches provide copious quantities of annotation available for transfer to new sequence, but adoption of terminology without a clear evidence trail can lead to the propagation of errors. Much of the annotated material in INSDC is the product of automated annotation pipelines, and may have a high rate of errors. It is necessary to find the evidence for a product statement in order to have confidence in the correctness of the prediction. This confidence is achieved by integrating evidence from many sources as described in the Methods of this work.

1.5 Aims of this work

Provide a basic annotation for the complete genome sequence of *S. scabies* 87.22 for the use of other researchers

Use this annotation to investigate the evidence in the genome for pathogenicity traits and regulatory mechanisms likely to control those traits.

From the genome sequence, assess the likely capacity of *S. scabies* 87.22 for complex natural product biosynthesis.

Develop methods for identifying and studying gene clusters likely to be involved in biosynthesis of complex natural products.

Objectives and hypotheses to meet aims

Indicators of known mechanisms of gene regulation will be found located in proximity to genes encoding known pathogenicity traits.

Such indicators of regulatory mechanisms may allow identification of gene targets for future laboratory investigations.

Identify potentially mobile regions by comparison with available sequences of related organisms.

Comparisons of the genome of *S. scabies* 87.22 with the non-pathogenic streptomycetes complete sequences to assist in identifying genetic material associated with pathogenicity and locating the boundaries of potentially mobile insertion/deletion regions.

It was expected that a very similar pathogenicity island would be found in the genome of *S. scabies* 87.22.

It was expected that biosynthetic genes for thaxtomins, extracellular esterase *estA*, necrogenic factor *nec1*, saponinase *tomA*, and biosynthetic genes for concanamycins would be found in the genome.

2 Methods for genome annotation

This chapter describes methods used for analysis and annotation of the complete genome sequence of *Streptomyces scabies* 87.22. Methods for detailed study of gene clusters likely to encode enzymes involved in biosynthesis of complex natural products are described in Chapter 3.

The primary aim of this annotation (as described above section **1.5**) was to provide a general basic prediction of the sequence features. This annotation is freely available under the Wellcome Trust Sanger Institute (WTSI) data release policy, which can be viewed here: **http://www.sanger.ac.uk/notices/release-policy.shtml**

The annotation is designed to be of use to the diverse research community likely to be interested in the sequence of *S. scabies* 87.22. This community includes those who are interested in *S. scabies* as a pathogen, and also those who are interested in the organism because it is a streptomycete.

The task of genomic annotation requires that a short statement is written of the likely purpose of the encoded protein, in the 'product' field of the coding sequence (CDS) feature. The content of this product statement is determined by evaluating and integrating data from a number of sources: sequence motifs, similarity searches and so on.

Note on presentation in this document

Names of software application, scripts and software commands are presented in `equal width` to make it clear that proper names of applications are being used. Sequence alignments are presented in `equal width font at a smaller size` for ease of layout.

2.1 Strains

See Table 2-1.

INSDC accession		organism	Details (module)	Publication(s)
nucleotide	protein			
DQ403252	ABD65960	*Streptomyces fungicidicus*	ErdD (M1)	Yin et al. 2006
AB211309	BAE98155	*Streptomyces lasaliensis*	Ecm6 (M1)	Watanabe *et al.* 2006)
CP000076	AAY91421	*Pseudomonas fluorescens*	PFL2147 (M3)	Gross *et al.* 2007
AL939115	CAB38517	*"Streptomyces coelicolor" str. A3(2)*	CdaII (M3)	Bentley *et al.* 2002; Hojati *et al.* 2002
AF210249	AAG02354	*Streptomyces verticillus*	BhmX (M1)	Du *et al.* 2000
AF255732	AAG27088	*Streptomyces acidiscabies str.* 84.104	TxtB	Healy *et al.* 2000
AF255732	AAG27087	*Streptomyces acidiscabies* str. 84.104	TxtA	Healy *et al.* 2000
AF047828	AAC80285	*Pseudomonas syringae* pv. *syringae*	SyrE (M2)	Guenzi *et al.* 1998
AL939105	CAB53322	*"Streptomyces coelicolor" str. A3(2)*	SCO0492/CdhH (M3)	Lautru *et al.* 2005
DQ118853	AAZ23077	*Streptomyces fradiae* NRRL18158	LptC (M1)	Miao *et al.* 2006
AAX31558	AAX31558	*Streptomyces roseosporus* NRRL 11379	DptBC (M1)	Miao *et al.* 2005
AY707081	AAW49318	*Streptomyces turgidiscabies* str. Car8	TxtA	Kers *et al.* 2005
AY707081	AAW49319	*Streptomyces turgidiscabies* str. Car8	TxtB	Kers *et al.* 2005
AY707080	.	*Streptomyces turgidiscabies* str. Car8	*necI* and others	Kers *et al.* 2005
AY707079	.	*Streptomyces turgidiscabies* str. Car8	*tomA* and others	Kers *et al.* 2005
EU119868	ABV21806	*Streptomyces scabies* 87-22	TxtR	Joshi *et al.* 2007
AF393159	AAL36833	*Streptomyces acidiscabies* 84104	TxtC	Healy *et al.* 2002
CAA63420	X92765	*Streptomyces diastatochromogenes* str. Tue20	EstA	Tesch *et al.* 1995
M57297	AAA26743	*Streptomyces scabies* str. FLI	EstA	Raymer *et al.* 1990
CAA33603	1AMU	*Brevibacillus brevis*	GrsA	Stachelhaus *et al.* 1995; Conti *et al.* 1997
.	2VSQ	*Bacillus subtilis*	SrfA-C	Tanovic *et al.* 2008
.	1MD9	*Bacillus subtilis*	DhbE	May *et al.* 2002
AJ300832	CAC83622	*Delftia acidovorans*	Fcs	Plaggenborg *et al.* 2001
NP_215056	CAB08976.1	*Mycobacterium tuberculosis* H37Rv	MenE	Cole *et al.* 1998
AF080217	Q9Z3R3	*Sinorhizobium meliloti* str. SU47 Rm1021	AcsA	Cai *et al.* 2000
AJ575648	CAE02600	*Streptomyces thioluteus*	AurE	He *et al.* 2003
P13129	1LCI	*Photinus pyralis*	luciferase	Conti *et al.* 1996
	1DNY	*Brevibacillus brevis*	TycC3	Weber et al. 2000
	1L5A	*Vibrio cholerae*	VibH	Keating et al. 2002
		Streptomyces coelicolor A3(2)	whole genome sequenc	Bentley *et al.* 2002
		Streptomyces avermitilis M4-4680	whole genome sequenc	Ikeda *et al.* 2003

Table 2-1Strains and organisms referred to in the text and used in sequence comparison studies.

2.2 Sequencing

A whole genome shotgun library was created in *Escherichia coli* and the first draft sequence was assembled from dye-terminator Sanger capillary reads on Applied Biosystems 3730 DNA analyzers at Wellcome Trust Sanger Institute. The sequence was finished by Kathy Seeger of WTSI pathogen sequence finishing team under the supervision of David Harris. The finishing team used the DYEnamic ET Terminator Cycle kit (Illustra TempliPhi Sequence Resolver Kit from GE Healthcare formerly Amersham Biosciences) to close gaps and resolve uncertainties. Finishing was done to a minimum of two strands covering each base with quality checking using GAP assembly software (Bonfield and Staden 1995). No independent plasmids or other extra-chromosomal DNA were observed in the genomic DNA preparation.

If there was any reason to be uncertain about a base call during annotation, such as a single nucleotide difference that changed the amino acid codon compared to closely related organisms, the assembly was checked on request. In all such cases the base call was found to be confirmed in multiple reads.

2.3 Automated pipeline design for preliminary annotation

An automated annotation pipeline was applied to the finished whole genome sequence. The sequence including transferred annotation was then inspected and edited by the annotator. The scripts comprising the automated annotation pipeline have been written and edited by programmers in the Pathogen Sequencing Unit at WTSI.

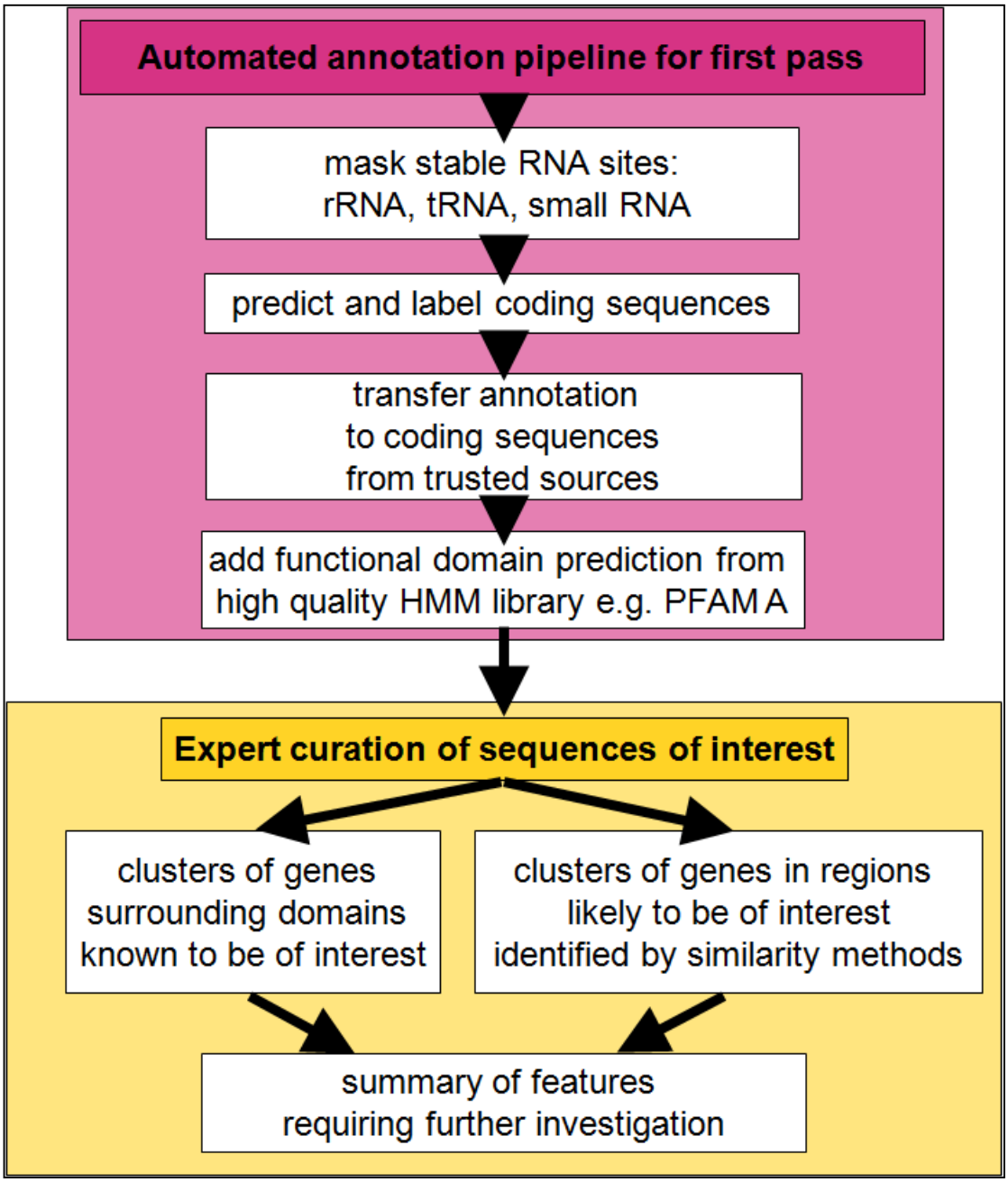


Figure 2-1 Summary of process for genome investigation. Automatable functions in magenta area, annotator's functions in yellow area.

2.3.1 Coding sequence prediction

Prokaryotic gene finder `Glimmer 3` (Salzberg and Delcher 2004) was used to predict coding sequences on the complete genome sequence. `Glimmer` is a prokaryotic CDS finder (Salzberg *et al.* 1998; Delcher *et al.* 1999; Salzberg and Delcher 2004).

2.3.2 Coding sequence numbering from SCAB00011 in tens

The coding sequences predicted in the genome of *Streptomyces scabies* 87.22 have been numbered from SCAB00011 to SCAB91071. These identifiers differ by tens, to allow room for additional coding sequences judged by the annotator during inspection to be false negatives by `Glimmer`, such that the next CDS after SCAB00011 is SCAB00021.

2.3.3 Import of annotation from trusted sources

Annotation was imported from the internal database in the Pathogen Sequencing Unit, WTSI. This is compiled from the complete annotators' version of all the microbial genomes sequenced so far at WTSI. Entries from this database were imported by script where a reciprocal `fasta` (Pearson and Lipman 1988) match was found over more than 80% of the length of the predicted CDS in *Streptomyces scabies* 87.22 with more than 30% identical residues. The annotator's version is the one created for genome analysis, which may have a great deal more information than the standardized version curated by INSDC.

2.4 Sequence visualisation

2.4.1 Artemis

`Artemis` (Mural 2000; Rutherford *et al.* 2000; Berriman and Rutherford 2003; Abbott *et al.* 2005; Carver *et al.* 2005) was used for sequence viewing and annotation; `fasta` (Pearson and Lipman 1988) searches were run from `Artemis` onsite at Wellcome Trust Sanger Institute (WTSI) using the pathogen sequencing unit cluster for visualization of alignments and INSDC accession data in `Artemis`'s Object Editor window. `Fasta` searches are used for this purpose because the `fasta` method attempts to find full-length matches, whereas the local alignment tools are more likely to pick up on matches to a dingle domain or region in a protein.

The annotation version has been released via strepdb **http://strepdb.streptomyces.org.uk/** upon completion in accordance with WTSI data release policy. Submission to INSDC requires much of the detailed annotation used in analysis (including /colour and /class qualifiers, and the status indicator ISS/IEA) to

be removed in order to provide a somewhat uniform annotation across research groups releasing sequence data. The complete genome sequence has been submitted and can be found by searching for the accession number FN554889.

Large genome mode

`Artemis` was run in large genome mode (10.1 M base pairs is exceptionally large for a bacterial genome) by passing the `java` virtual machine a parameter `-Xmx256M.` This authorizes the `java` virtual machine (upon which `Artemis` is run) to use a larger memory allowance. This was done via the command line or in `Microsoft Windows` via a dedicated shortcut. An example command for the shortcut (periods used to indicate path to the file) could be something like:

C:\...\java.exe -Xmx256M -jar "C:\...\artemis_v9.jar"

Useful Artemis features

Double click on a feature to centre the 5' end.

Zoom in and out from the sequence using side bars.

The annotation was written as a .tab file in `Artemis`. Coding sequence features were edited using Artemis Feature Edit (Figure 2-2).

Extensive use was made of Artemis's Object Editor which allows instant visualization of the extent of `fasta` matches against `UNIPROT` along with any conserved domains from `Pfam` (Finn *et al.* 2006) so that it is apparent whether matches are just to conserved domains or more useful overall conserved architecture.

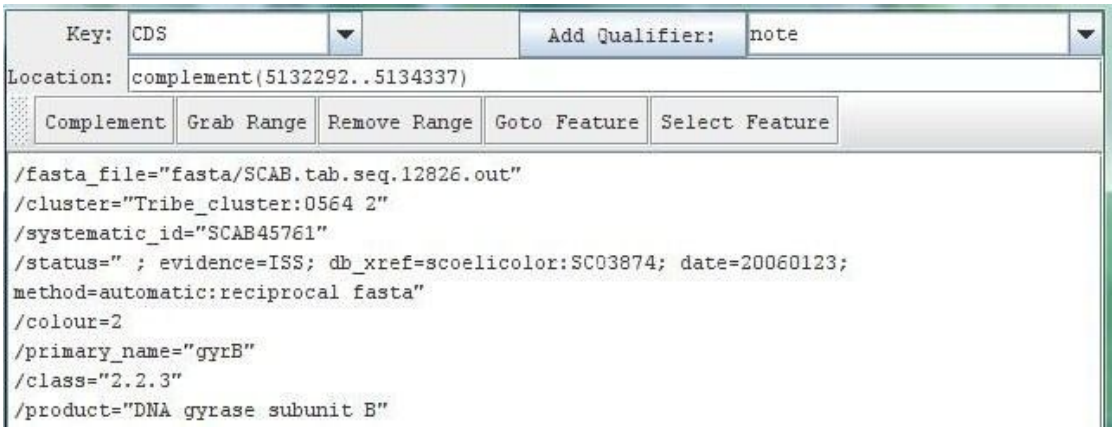


Figure 2-2 **Artemis Feature Editor showing coding sequence *gyrB*.** Key is the kind of feature being studies; qualifiers are the fields recognized by Artemis and preceded by /.

2.4.2 Artemis Comparison Tool (ACT)

Artemis Comparison Tool was used to visualize the extent of matches between sequence in "*S. coelicolor*" A3(2), *S. avermitilis* MA-4680 and *S. scabies* 87.22, and comparisons with sequences from other organisms as relevant for example through study of similar gene clusters as described in Chapter 3. Comparison files for use in ACT were generated using the same blast implementation as described below with appropriate options for the particular comparison.

2.5 Data curation methods

Several datasets as described below were integrated into the annotator's version of the genome annotation (SCAB.tab). All available lines of evidence were used together to decide on the likelihood of the coding sequence being genuine and the appropriate product line for the coding sequence feature. Similarity searches were used where possible to identify a characterised protein with identical domains and a high levels sequence similarity. If this was not possible the evidence of domains, motifs and similarity searches was used to suggest a function, such as by downgrading the specificity of a protein's gene function to the broad function of the protein family. "Putative" was used to indicate a guess, but since almost all annotation is guesswork this is not very useful.

Citations were added to the annotation using the /citation qualifier to add the PubMed **http://www.ncbi.nlm.nih.gov/pubmed/** accession of relevant literature, for example publications with related sequence or characterisation data.

The Kyoto encyclopedia of gene and genomes (KEGG) pathways (Ogata *et al.* 1999) were used together with Enzyme Commission classification website **http://www.chem.qmul.ac.uk/iubmb/enzyme/search.html** to identify enzymes predicted to be encoded by coding sequences occur in close proximity. Domain and similarity-based information about the coding sequences was compared with enzyme commission reaction labels.

2.5.1 Version control

As layers of annotation were added to the main file (SCAB.tab) from each application used, a version was archived (in the genome project directory on the PSU file space at WTSI) with an informative name. During the lengthy phase of analysis following annotation new version files were created at regular interval labelled with an indicator of the date (YYYY-MM-[name] is easily machine sorted in correct order.)

2.5.2 Stable RNA predictions

Ribosomal RNA operons were marked by generating sequence features from an example operon. An rRNA operon sequence from *Streptomyces coelicolor* A3(2) (Bentley *et al.* 2002) was formatted as the database for a `blastn` search. From the `blastn` results, sequence decoration was generated for integration into the main annotation file.

Transfer RNA (tRNA) sequences can be found from secondary structure prediction models, but not all of these give identical results. `tRNAscan` (Lowe and Eddy 1997), and `Rfam` (Gardner *et al.* 2009) were both used to predict tRNA locations. The set used in the annotation is the union of both prediction sets, assuming that it is more likely that variant tRNAs would be missed than tRNA-like features wrongly marked. Several other stable RNA features were also identified using the version of `Rfam` available at the time (Griffiths-Jones *et al.* 2003).

2.5.3 Curation of CDS prediction

6849 coding sequence features were inspected and edited. If both frame plot (0) and correlation score (2.5.3.2) met the expected values for this genome, but no significant matches (E>0.01) were apparent from `fasta` vs UNIPROT (Uniprot 2007), these

coding sequences were indicated in `Artemis` with colour=8 and product line 'hypothetical protein'. 'Conserved hypothetical protein' is used as the product line for coding sequences where significant (E<0.01) matches are found in UNIPROT, but no characterisation studies have been located sufficient to confidently infer function. Conserved hypothetical protein predictions are marked with colour 10.

Coding sequences identified in the `Glimmer 3` prediction and to which preliminary annotation was added by the automated annotation pipeline were given an indicator IEA (Inferred by Electronic Annotation) in the /status qualifier. During curation this IEA indicator was replaced with ISS (Inferred from Sequence Similarity), in order to provide a record of which CDS have been inspected by an annotator. IEA and ISS are a subset of the GO evidence codes for annotators and the new release (Rogers and Ben-Hur 2009) could be used in future work.

The set of coding sequences described in this work was been released under the WTSI data release policy as soon as completed. This set is used for the descriptive statistics are based in the summaries presented within this work. Any set of coding feature predictions should be regarded as provisional.

2.5.3.1 FRAME analysis

A plot of third codon G+C content, known as FRAME (Bibb *et al.* 1984), helps to indicate the presence of a coding sequence in organisms with genomic G+C content deviating from 50% (Bibb *et al.* 1984). Because of redundancy in the third position ('wobble position') of amino acids codons, the G+C content of these positions is likely to have resulted from mutation bias (Wright and Bibb 1992) towards the organism's preference in G+C content.

The N-terminal extents of coding sequences were adjusted with reference to FRAME plot. Missing coding sequences in the automated CDS prediction have been added where the plot made it clear, and false positives removed.

2.5.3.2 Correlation scores

Correlation score (Rutherford *et al.* 2000) is a metric for the correlation between amino acid composition of globular proteins in TREMBL and the translation in each

reading frame. Correlation scores greater than 52 are fairly likely to indicate the presence of a coding sequence (S. D. Bentley pers. comm.) Scores were used as one line of evidence indicating the presence of a coding sequence as part of adjudication of doubtful coding sequences.

2.5.3.3 Ribosome- binding sites

Identification of the ribosome-binding site can define the N-terminal extent of the CDS. Considerable variation is apparent in the regions immediately upstream of coding sequences in streptomycete genetic material (Strohl 1992), but a ribosome binding site of Shine-Delgarno () type having the form 5'-GGAGG-3' may be found in some cases. Other researchers have looked for four out of five of the above consensus (Choulet *et al.* 2006a). MEDstart (Zhu *et al.* 2004) could be useful for future investigations.

2.5.3.4 Tblastx overlay

A set of sequence decorations was constructed from `tblastx` (Altschul *et al.* 1990; Altschul *et al.* 1997) against UNIPROT (Uniprot 2007). This method highlighted a false negative from Glimmer 3 in at least one case, the *nec1* sequence (see further **1.3.4**).

2.5.3.5 Proteomics data

A set of preliminary proteomics data gathered by R. Loria and colleagues was visualized in Artemis and it was intended that this would be useful for confirming doubtful CDSs. Some of the features generated from this set fell on the complementary strand to the likely coding sequences as judge by FRAME, correlation score and similarity search so it was not useful for this purpose.

2.5.3.6 Pseudogenes

Where frameshifted pseudogenes were identified by similarity searches, several coding sequence features were merged to create a single pseudogene feature using `Artemis Edit> Selected feature(s)> Merge.` Inframe STOP codons were excluded from the pseudogene CDS in order to make the most useful kind of feature for searches by future users of the annotation. A /pseudo qualifier was added

to the coding sequence feature, and a /note qualifier indicating the reason for assuming the coding sequence is not functional.

2.5.4 Colour and class qualifiers

A colour code in use amongst annotators at WTSI (Appenidx A) was used to broadly classify coding sequences. This colour code is an aid to the annotator because it helps mark out genes of related function at a glance, so transport systems for example stand out. The `/colour` qualifier was used to store the digits associated with each colour, and `Artemis` colours the coding sequences according to it. The presentation of such data in visual form has clear advantages for allowing "reading" of a section of sequence.

A functional classification system under the `/class` qualifier was also used (Appendix A) to classify coding sequences for text-based interrogation of the annotation file during analysis.

2.5.5 Domain and motif searches

Domain and motif searches were used to back up similarity-based inference and to suggest function for hypothetical proteins. This allows consideration of the modular architecture of proteins – all of the domains must be conserved to give confidence that a new sequence will encode an enzyme with the same function. Inferring from the presence of functional domains increases sensitivity because only the evolutionarily conserved regions of the protein are taken into account (Parkhill 2002).

Modular architecture of proteins has been noted as a source of error in annotations (Galperin and Koonin 1998; Brenner 1999). Conserved function should only be inferred if a check has been made to ensure that all if the same conserved domains are present in multienzymes proteins.

`Interpro` including `Pfam`

Sequence decoration was generated from matches to conserved domains and motifs in `InterproScan` (Zdobnov and Apweiler 2001; Quevillon *et al.* 2005) databases as follows: `TMHMM` for transmembrane protein prediction (Krogh *et al.* 2001), `HMMPFAM` for the `Pfam` library of conserved protein domains (Bateman *et al.* 1999;

Bateman *et al.* 2000; Bateman *et al.* 2002; Bateman *et al.* 2004; Finn *et al.* 2006), PROSITE motifs (Bairoch 1992), and SignalPHMM to predict protein secretion signals (Nielsen *et al.* 1999) as supporting evidence for functional annotation.

The Pfam database (Finn *et al.* 2006) has two parts – a curated part matching nearly 9000 protein families and a supplement, Pfam-B containing a large number of small families without curation. Domain architecture of predicted proteins was checked using the "bead on a string" graphics provided on the Pfam website to avoid inappropriate transfer of annotation. Sequence features representing the conserved domains of predicted proteins were imported into Artemis as sequence decorations using the ipro2tab.pl script pathogen sequencing unit computing cluster at WTSI. When close examination of a coding sequence features was necessary, Pfam conserved domain searches were rerun to find fragmentary and sub-threshold matches in case they were indicative of a protein family for the hypothetical protein.

Structural features and sub-cellular location

Predicted proteins with one transmembrane helix prediction from TMHMM were annotated as "membrane protein"; with several transmembrane helices, "integral membrane protein" unless there was evidence from similarity to a characterised transport protein or conserved Pfam domain architecture to infer transport function. Kyte-Doolittle hydropathy plots (Kyte and Doolittle 1982) were visualized in Artemis to check hydrophobic regions in protein predictions (window size 9 for hydrophobic globular proteins, window size 19 for transmembrane helices). A new development since the automated predictions were added to the genome annotation is sub-cellular location prediction based on conserved protein domains which could be useful in future work (Guo *et al.* 2006).

Where hydrophobic N-terminal secretion signals were evident from SignalPHMM (Nielsen *et al.* 1999) results with no other evidence of function the predicted protein was annotated as 'secreted protein'. In case of uncertainty, location prediction was checked using PSORTb web server (Gardy *et al.* 2005) in Gram positive mode. PSORTb performs well in comparison of software tools for prokaryotic sub-cellular location prediction (Gardy and Brinkman 2006). LocateP looks like a good candidate for sub-cellular location prediction in future investigations (Zhou *et al.* 2008).

2.5.6 Detection of mobile elements in sequence data

Detection of mobile elements from nucleotide sequence can draw on several approaches: regions of difference from other genomes as described below (**3.1.1)**; conserved domains associated with mobility as described in **1.1.5**; and sequence composition features. Comparison with related genomes can reveal regions where insertions or deletions have resulted in gene acquisition or loss. Signals from compositional bias and conserved domains might also be detected, and these will depend on the kind of element and the origin of the donor. A G+C signal might be observed if there is a significant difference in G+C composition between donor and recipient. However, if the element has been transferred from a member of the same species or genus, compositional bias signals will be much less evident.

Conserved domains known to catalyse mobility can be identified using the various kinds of conserved domain approach. Prophages and plasmids have characteristic domains which can assist in identification of those kinds of features, as discussed above (**1.1.5)**. The presence of domains with integrase function is particularly significant since this is required for excision of a mobile element and may indicate the presence of an element which is capable of effecting transfer.

Identification of repeats might serve as confirming evidence once a potential mobile element is identified. Repeats can be very small so because they can be common by chance alone in genomes, identification of repeats alone (for example using reputer (Kurtz and Schleiermacher 1999)) was not a useful strategy for first pass identification of mobility. Because tRNA genes are the most common target for site-specific integration, study of the regions flanking these sequences often reveals mobile elements, so alien_hunter (Vernikos and Parkhill 2006; Langille *et al.* 2008) was be used generate a list of regions of possibly mobility. `Alien_hunter` integrates information about conserved domains adjacent to tRNA loci with nucleotide deviation signatures.

Compositional deviation can indicate mobile elements, especially if the donor organism has a significantly different make-up than the recipient. Deviation in G+C content signal is commonly attributed to horizontal transfer between organism with widely divergent G+C content, but several researchers have found this badly

supported by more careful discrimination of horizontally transferred genes (Koski *et al.* 2001), and G+C signatures can also indicate genes carried in phage genomes (Daubin *et al.* 2003) so compositional deviation has not been used for identification of regions for study in this work.

Deviation in dinucleotide (Karlin signature) (Karlin *et al.* 1997) and higher order nucleotide groups also can be used (Pride *et al.* 2003) (Vernikos and Parkhill 2006), to detect changes in compositional bias. Changes in apparent codon usage may also indicate transferred material (Zhang and Zhang 2004) and this is unsurprising if both lower order (dinucleotide) and higher order signatures show the same effect - codon preference is the same as trinucleotide usage.

G+C skew is another kind of compositional deviation which may be useful in the identification of mobile elements. There appears to be a preference for incorporation of guanine in the leading, and cytosine in the lagging, strand of replication in many bacterial genomes (Gruss and Michel 2001). Change of sign in G+C skew hence can be used to identify the origin of replication in bacterial genomes (McLean *et al.* 1998), and deviation from the expected pattern of G+C skew could indicate integration of material which has spent a long time in different position on a chromosome. The G+C skew graph can be visualized using Artemis (Rutherford *et al.* 2000).

2.6 Phylogenetic methods

`Clustalx` (Thompson *et al.* 2002), a Microsoft Windows application to run the `clustal` group of programmes, and `muscle` (Edgar 2004) were used to build pairwise and multiple sequence alignments, with distance matrix BLOSUM62 (Thompson *et al.* 1994) for amino acid sequences of length 85-300 residues. Alignments were edited where considerations from models of the protein structure indicated constraints were likely to act, and to ensure sequenced compared were of equivalent length. Applications from `Phylip` (Felsenstein 2008) were used for testing the fit of models of evolution to aligned sequences (`modeltest`), calculation of distance matrices (`protdist`), generation of 1000 bootstrap replicate sets

(seqboot), and extended-majority rule evaluation of those replicate tree sets (consense).

2.6.1 Global pairwise alignment of coding sequences

Global pairwise alignments for comparison of individual coding sequences were performed using needle, an implementation of the Needleman-Wunsch alignment (Needleman and Wunsch 1970) algorithm, onsite at WTSI. Needle settings for amino acid sequences length 85-300 residues: matrix=EBLOSUM62, -gapopen=12, -gapextend=2; for nucleotide sequences matrix=EDNAFULL, -gapopen=16, -gapextend=4.

2.6.2 Treebuilding

Neighbour joining trees (Saitou and Nei 1987) have been constructed because they are computationally efficient and statistically valid (Gascuel and Steel 2006), and easily to construct using the software packages available. Phylogeny is a complex field deserving of in-depth study and a competent researcher might consider using a modular system such as Mesquite (Maddison and Maddison 2009). Trees were constructed using two outgroups to root the tree where possible, and were visualized as unrooted or radial trees using Phylodraw (Choi, J. H. *et al.* 2000).

2.6.3 Phylogenetic workflows

Protein sequence studies

Import protein alignment to Phylip

First pass: Modeltest > Protdist > Neighbor

Full study: Seqboot > Protdist > Neighbor > Consense

2.7 Additional resources

2.7.1 Basic local alignment search tool (`blast`)

An implementation of the `blast` suite of algorithms (Altschul *et al.* 1990) was used extensively via the `bigger_blast.pl` wrapper script on computer clusters at Wellcome Trust Sanger Institute. These were used for comparing whole genome alignments; for checking the automated coding sequence prediction against known protein sequences; and for comparing gene clusters expected to encode biosynthetic pathways for complex natural products.

2.7.2 Dotter

Dotter (Sonnhammer and Durbin 1995) was used for rapid direct visualization of high scoring pairs in matches between nucleotide and protein sequences. Repeated elements are made visible by comparing a sequence against itself, and the extent of primary protein sequence identity is revealed by comparing two against each other.

2.7.3 Clustering methods

Tribe clustering

A clustering metric, `Tribe MCL` (Enright *et al.* 2002), was used within the genome to help in identifying certain families of coding sequences likely to be related to each other, for example insertion sequences. An attempt was made to standardise annotation across the clusters recovered using `Tribe`, using `Artemis` (Rutherford *et al.* 2000) to collect the presumably related sequences. Output from this script was imported as sequence features with a /cluster qualifier of the form shown below. The four figure digit is the cluster number, and the number of coding sequences within the cluster follows. The cluster can be recovered whilst visualizing the annotation by using `Artemis`'s Feature Selector to select CDS features containing the qualifier "cluster" with the text of the cluster number.

Example: `/cluster="Tribe_cluster:0564 2"`

Orthologue clustering

Orthologues - proteins performing the same function and most closely related to a common ancestral protein – within the three *Streptomyces* whole genome sequences were identified for annotation purposes by `fasta` (Pearson and Lipman 1988) matches >70% amino acid sequence identity with conserved domain architecture. An implementation of the `Orthomcl` (Li, L. *et al.* 2003; Chen *et al.* 2006) clustering algorithm was also used. A subsequent published study (Chater and Chandra 2006) has determined tables of orthologues between the three available streptomycete genomes and there would be no obvious reason not to use that work.

2.7.4 TTA codon scripts

A script for locating TTA codons was kindly shared by the authors who investigated TTA codons in *"Streptomyces coelicolor"* A3(2) (Li, W. C. *et al.* 2007). The locations of TTA codons reported by this script were imported into the annotation file using a perl script to construct a /note qualifier describing the position(s) of TTA codons.

2.7.5 Promoter pattern searches

`Artemis` (Rutherford *et al.* 2000) was used to generate sequence decorations illustrating sequence matching patterns. "Mark From Pattern" was selected from the "Create" menu, and a nucleotide sequence entered, using the IUB codes for incompletely specified bases (Cornish-Bowden 1985). The files generated by `Artemis` describing the matches were saved and transferred into Microsoft Excel worksheets for examination. The search pattern was repeated with greater or lesser redundancy until to find the largest possible set of sequences which matched both strands.

In this work a brief study was undertaken beginning with patterns from previously-identified iron box sequences (Gunter *et al.* 1993; Wisedchaisri *et al.* 2004; Yellaboina *et al.* 2004a). Regulon prediction is an exciting field and a software tool such as PREDetector (Yellaboina *et al.* 2004b; Hiard *et al.* 2007) have been used with promising results (Rigali *et al.* 2004; Rigali *et al.* 2006; Hiard *et al.* 2007) and show

potential for beginning to make good use of genome sequence data for investigations into the biology of organisms.

3 Methods for in-depth study of gene clusters

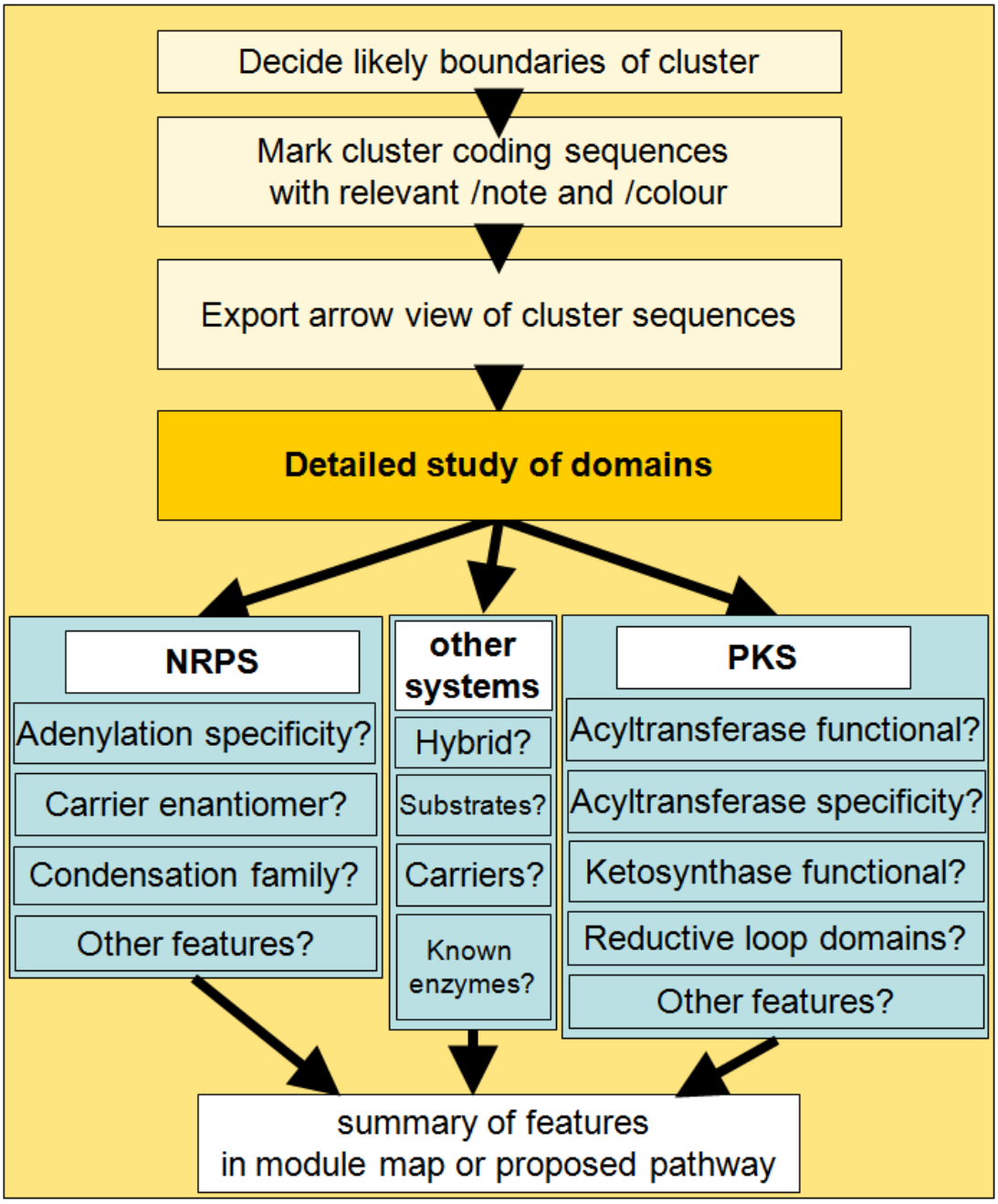


Figure 3-1 Overview of process for in-depth annotation of gene clusters suspected to encode biosynthetic enzymes for complex natural products. This flow chart represents a serious of laborious tasks which are not yet clearly transferable to automated processes.

3.1 Identifying probable complex product gene clusters

The clusters identified by these methods as *likely* to be involved in complex product biosynthesis are not proven to do so. This work identifies clusters to provide targets for further investigation on the basis of conserved domains likely to be involved in

such biosynthesis, or by sequence similarity to known natural product biosynthetic enzymes. The methods described here were developed during this project and were used by the author in a subsequent investigation to study a novel non-ribosomal peptide synthetase (NRPS) system found in the genome of *Pseudomonas fluorescens* SBW25 (Silby *et al.* 2009).

Clusters identified in this work have been numbered for convenience by type beginning at base number 1, for example *nrps1* describes the genes proposed to encode the first non-ribosomal peptide synthetase system identified, *nrps2* the second and so on.

It is almost inevitable that the reported set of coding sequences for complex product gene clusters in this work is incomplete. Gene clusters involved in complex product biosynthesis can be as small as one gene, as is the case for geosmin (Cane and Watt 2003; Gust *et al.* 2003; Jiang *et al.* 2006). Hence, some coding sequences annotated *hypothetical protein* and *conserved hypothetical protein* will be found in future investigations to encode complex product biosynthesis amongst other functions which are not yet known.

The most well-studied systems for biosynthesis of complex natural products are the non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) systems. Several other kinds exist such as those involved in biosynthesis of desferrioxamine siderophores (Challis 2005) and the terpene cyclases involved in production of aromatic compound geosmin (Jiang *et al.* 2006) and 2-methylisoborneol (Komatsu *et al.* 2008). The genes involved in biosynthesis of products other than NRPS and PKS systems have to be identified by a combination of similarity searches and identification of conserved domains. It is necessary for the annotator to know the identity of complex natural products for which the genes have been identified and the conserved domains associated with them. Hence a typical study on a non-NRPS/PKS system involved finding a characterised gene cluster with similarities, and comparing the results of conserved domain searches on the candidate and the characterised cluster. The study of NRPS and PKS systems is similar, but it is easier to describe the set of characteristic domains involved because there is a relatively small number of domains that re-occur in various combinations to produce the vast diversity of products.

Gene clusters likely to encode enzymes involved in biosynthesis of complex natural products were identified by the presence of conserved domains involved in such pathways, or by results of similarity searches. The details of these methods are explained below.

3.1.1 Is this cluster the same as that cluster?

Comparisons have been made between the three streptomycete genomes ("*S. coelicolor*" A3(2) (Bentley *et al.* 2002), *S. avermitilis* MA-4680 (Ikeda *et al.* 2003), and *S. scabies* 87.22) to determine as far as possible which clusters are conserved across the three and which are only found in *S. scabies* 87.22. Artemis Comparison Tool (Carver *et al.* 2005) was used to visualize `tblastx` matches between gene clusters. The translated blast algorithm indicates where the amino acids of the primary protein sequence are identical.

Matches with high scoring pairs in `tblastx` between pairs of genomes were examined in ACT to determine whether the same conserved protein domains were present using `Interpro` libraries as previously described in **2.5.5 Domain and motif searches**. Clusters expected to encode biosynthetic pathways for identical products have been judged on the criteria described below **3.3.1 The same, or different, product?** A `.tab` file for Artemis (Rutherford *et al.* 2000) was compiled for close comparison, where the active site or conserved elements of enzymatic domains were marked for easy comparison. Each set of domains was marked with features in a particular colour to make it easy to pick out the confirmed domains. The `.dna` file associated with this detailed annotation of the cluster, and the `.tab` file containing the annotation, have been appended to the electronic indices.

3.1.2 Conserved domains as clues to biosynthesis

NRPS (non-ribosomal peptide synthetase) and PKS (polyketide synthase) systems have characteristic sets of domains. `Pfam` (Finn *et al.* 2006) was used to identify conserved domains in the complete genome sequence as described in **2.5.4.1 Interpro**. (Pfam A domains have identifying keywords of the form PF+digits.)

In NRPS systems: adenylation domains matching PF00501; peptidyl carrier protein domains matching PF00550; condensation domains matching PF00668; and

thioesterase domains PF00975 are expected. In PKS systems the domains expected are: acyltransferase PF00698; acyl carrier protein also matching PF00550; ketosynthase (N-terminal part PF00109, C term PF02801); and reductive loop domains for which the `Pfam` A models are not completely informative, ketoreductase, dehydratase and enoylreductase. Gene clusters likely to encode enzymes for biosynthesis of hydroxamate siderophores, for example the biosynthesis cluster for desferrioxamines (Barona-Gomez *et al.* 2004; Kadi *et al.* 2007) can be recognised by PF04183.

3.1.3 Similarity as a clue to biosynthesis

Highly ranked blast matches were found in some places during curation of the genome to genetic material previously sequenced as a result of investigation into known complex product biosynthesis clusters. Gene clusters for biosynthesis of complex natural products have been frequently published which include hypothetical proteins, the role of which is not known. Even unsubstantiated predicted proteins - previously sequenced as part of a biosynthetic gene cluster, but having unknown function - could be a clue to the presence of a biosynthesis cluster. Such references have been indicated by use of the `/citation` qualifier in `Artemis` (Rutherford *et al.* 2000; Berriman and Rutherford 2003) as described in **2.5 Data curation methods**.

Text searches were performed on the list of highly ranked blast matches for strings such as “biosynthesis”, in order to rapidly identify matches to genetic material previously sequenced as pat of another complex product biosynthesis gene cluster.

3.1.4 Boundaries of gene clusters

Evidence for possible involvement in complex product biosynthesis was noted during the first-pass annotation of the genome and locations of likely clusters of genes were noted. A second phase of detailed annotation was undertaken on around 250 coding sequences in these clusters. This in-depth second pass annotation involved inferring a boundary for the gene cluster and investigating the suggested function of predicted enzyme domains where possible.

Boundaries for gene clusters have been suggested after considering a number of lines of evidence. Co-transcription in operons was inferred where closely arranged or

overlapping coding sequences were found. It was then inferred that the boundary of the gene cluster would be likely to lie beyond the end of the apparent operon.

Some predicted proteins other than those likely to be involved in biosynthesis have characteristic domains. For example, regulators of gene clusters such as those predicted to carry the conserved bacterial transcriptional activator domain PF03704 (Yeats *et al.* 2003) were inferred to be part of a complex product biosynthetic gene cluster where found at the border of one. The domain matching PF03704 in characterised regulator AfsR has been found to independently direct transcription of the complex natural product actinorhodin (Horinouchi *et al.* 1990), and several proteins of this kind were found associated with complex product biosynthesis cluster identified in this work. Other coding sequences such as two component sensor/regulator pairs, frequently found in the model streptomycete "*S. coelicolor*" A3(2) (Hutchings *et al.* 2004) could also be inferred to be part of a gene cluster, though not encoding enzymes likely to be directly involved in biosynthesis.

The boundary of a complex product biosynthetic gene cluster might also be inferred by a clear function being apparent for neighbouring genes – if clearly part of core metabolic functions conserved across organisms, for example.

3.1.5 Colour qualifier

Clusters of genes inferred as likely to encode genes involved in biosynthesis of complex natural products have been marked in the annotator's version (available on request from WTSI PSU) of the annotation with colour 15, a brownish red. For more information on the use of the colour qualifier see **2.5.3 Colour and class qualifiers** and **Appendix A Classification and colour scheme for `Artemis`.** A `/note` qualifier has been added to the CDS feature for coding sequences suggested to be inside the cluster.

3.1.6 Export arrow view to illustration

A summary of each gene cluster studied in depth has been illustrated by exporting the cluster coding sequences from `Artemis` to form an arrow view illustration. The gene cluster was highlighted and a subsequence written out using `Edit> Subsequence and features`. The tab and sequence files were saved separately in `.embl` format

and an arbitrary descriptive header e.g. >nrps1 added. The tab and sequence files were then concatenated together using a text editor (nedit http://www.nedit.org/ or Microsoft Notepad). All annotation except the coding sequence feature and its cognate /product field are removed in Artemis.

Vector NTI Viewer

The combined file in .embl format (containing the coding sequences from the tab file and the DNA sequence of the gene cluster) was then opened in Invitrogen Vector NTI Viewer, a molecule viewer available via free download from the Invitrogen website http://www.invitrogen.com/site/us/en/home.html. Vector NTI Viewer is part of the Vector NTI package (Lu and Moriyama 2004).

Viewing options in Vector NTI Viewer

In Vector NTI Viewer: Display options> Picture type> prefer linear changes the display to linear view which is more representative. Also in Display Options window, restriction site mapping was removed to reduce clutter on the image.

To compose the arrow view of the gene cluster, an image is exported from Vector NTI Viewer to Microsoft Powerpoint. Vector NTI Viewer's "Camera" is used to export the cluster image to clipboard. When pasted into Microsoft Powerpoint, the elements of this image can be ungrouped to edit them.

3.2 Domains: detailed annotation

Once the boundaries of a gene cluster had been decided, the predicted enzymes were studied more closely. The aim of this level of annotation was to produce a detailed list of the enzymatic domains likely to be involved in biosynthesis of a complex natural product. Inevitably the level of detail in these studies is not the same as the level of detail possible when only one protein or protein family is being surveyed. Several recent advances are mentioned in the text below which will make this kind of work easier and better-quantified in future investigations such as Swiss model pipeline and resources (Kiefer *et al.* 2009).

3.2.1 Module map

After detailed annotation (as described below), a module map was composed to summarize functional domains predicted. This module map is presented with the arrow view of the proposed gene cluster as a summary figure at the head of each results section. Domains were identified using Pfam's A library (Finn *et al.* 2006) and in relevant literature described below. In order to comment on likely function, active site residues were checked for critical residues involved in catalysis, where known and where not included in Pfam models. Coding sequences without Pfam A domains were rerun against Pfam A and B to check for fragmentary, subthreshold, and uncharacterised domain matches as described in **2.5.4.1 Interpro**.

3.2.2 Nonribosomal peptide synthetase (NRPS) domains

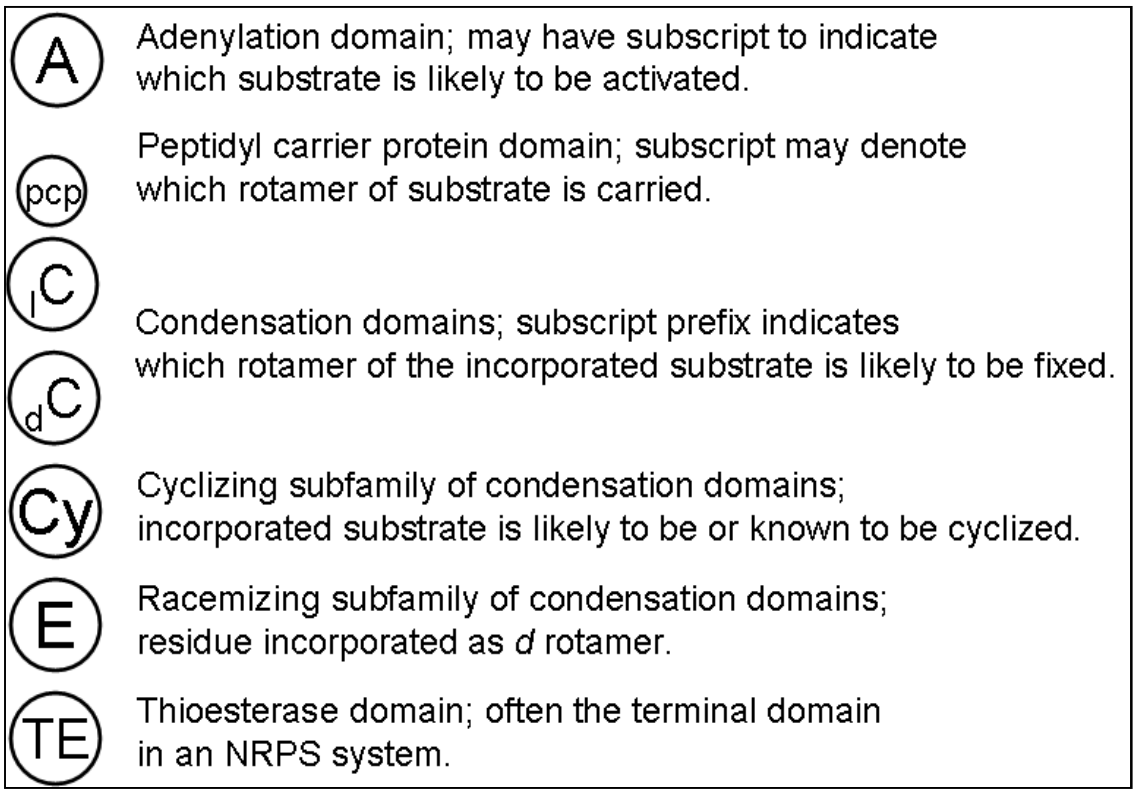


Figure 3-2 Key to NRPS module maps. Circles represent domains identified in an NRPS system, letters indicate the nature of the domains identified. Domains with a cross through them have been identified as having homology with known biosynthetic domains but lack critical residues in the active site and thus are judged as likely to be inactive.

3.2.2.1 Adenylation domains

Domains matching PF00501 in *S. scabies* 87.22 were aligned with reference sequences selected from public databanks.

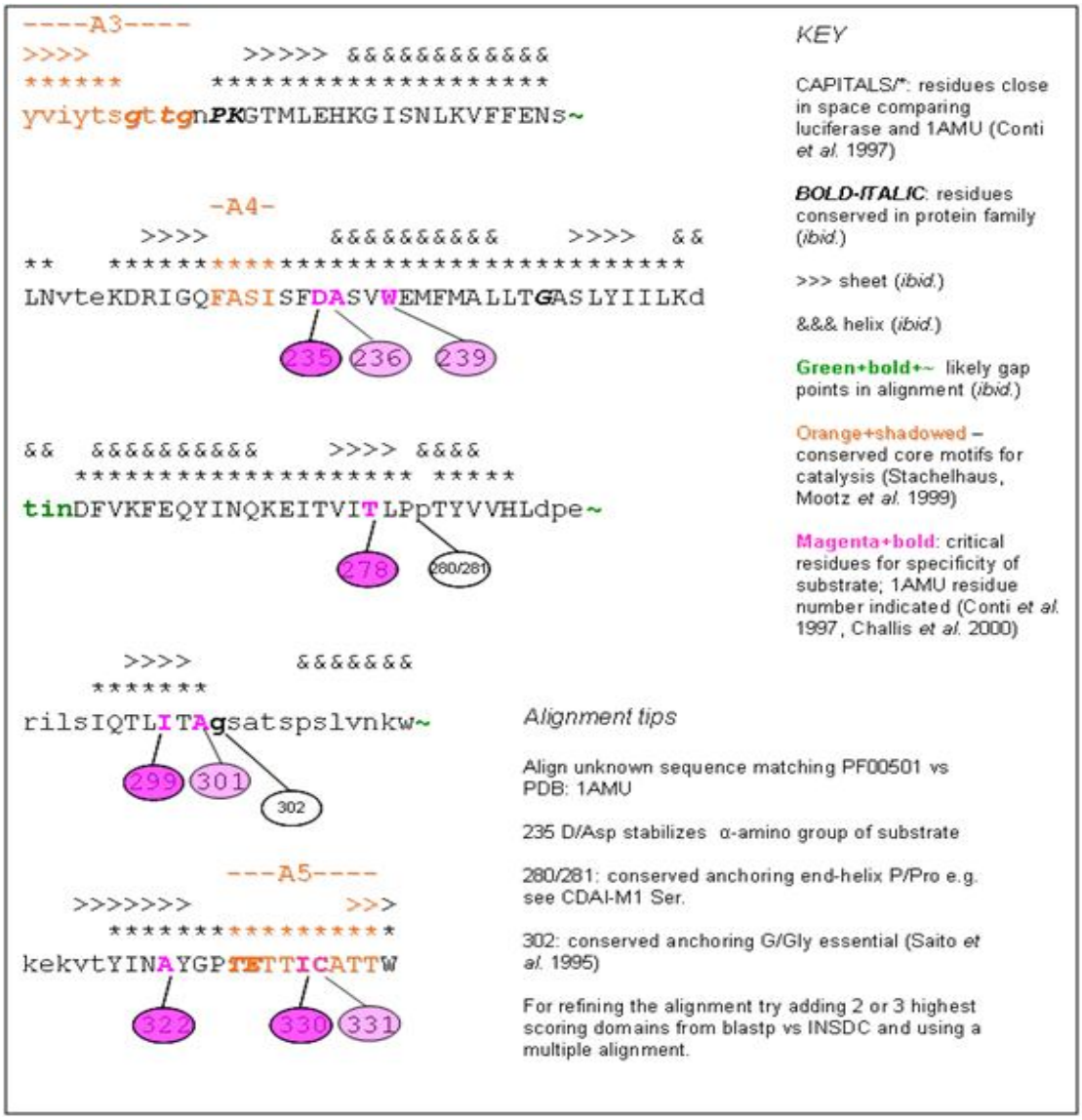


Figure 3-3 Alignment guide for adenylation domain alignments using 1AMU (Stachelhaus and Marahiel 1995) to identify critical residues for use of the predictive model (Challis *et al.* 2000)

Reference amino acid adenylation domain

The reference used for amino acid adenylation is the phenylalanine-activating domain of from the gramicidin S enzyme GrsA (PDB: 1AMU) in *Brevibacillus brevis* (Stachelhaus and Marahiel 1995; Conti *et al.* 1997). Comparison between this model and that of another adenylate-forming enzyme showed certain residues conserved across the family (Conti *et al.* 1997). This model was used by several groups of

researchers investigating the structural basis of specificity (Stachelhaus *et al.* 1999; Challis *et al.* 2000) and to generalize about the roles of certain residues in substrate selection (Challis *et al.* 2000).

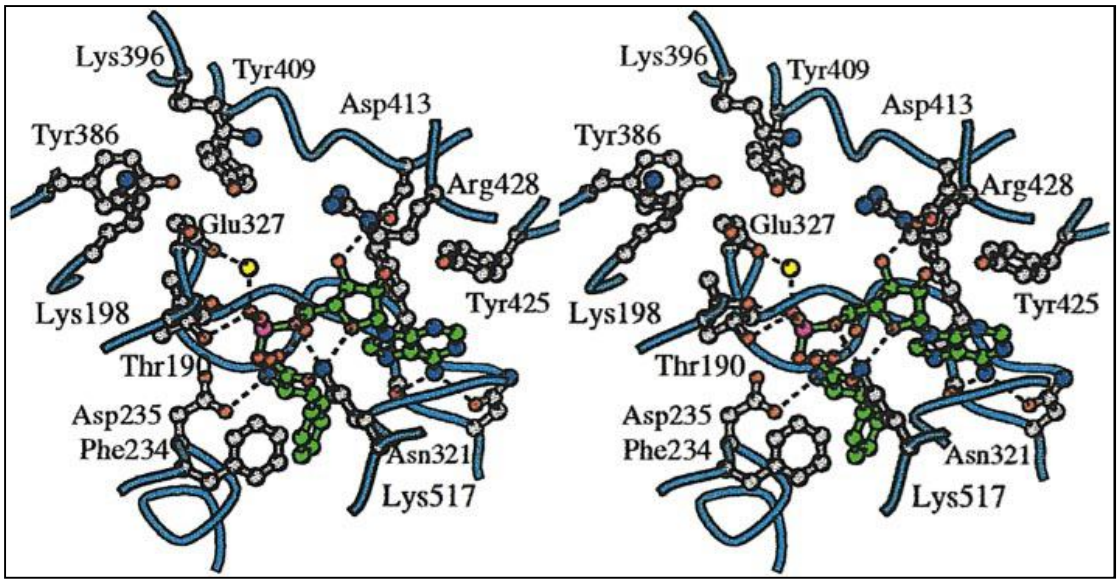


Figure 3-4 Binding pocket of 1AMU. From Conti, Stachehaus et al 1997: Stereo diagram showing the interactions made by the phenylalanine substrate and AMP (both in green) with the binding pocket of the adenylation domain. Numbers refer to the residue in this sequence 1AMU. The position of a possible magnesium ion is shown in yellow and potential hydrogen bonding interactions are indicated by dotted lines.

Domains with conserved sequences typical of thoas activating amino acid substrates have been studied using the structure-based predictive model (Challis *et al.* 2000) based on alignment with 1AMU (Stachelhaus and Marahiel 1995). This amino acid activating subset appears to form a clade amongst the sequences matching PF00501 (Figure 3-7) which provides a quick check preceding examination of the sequence for conserved residues. The critical residues providing the specificity conferring contacts in the domain have been identified using alignments of a short (~200 bp) protein sequence carrying the critical residues and bounded by sequences reported as conserved in the protein family (Conti *et al.* 1997) as illustrated in (**Error! Reference source not found.**).

Specificity of amino acid activating domains

Where the critical residues thus identified are identical to a set conferring known specificity, for example as found in the data table upon which the predictive model is based (Challis *et al.* 2000), it has been assumed that the adenylation domain has this

specificity. Where the critical residues in the binding pocket were not identical to those of a domain of known specificity, additional methods have been used to suggest a possible specificity for the binding pocket.

The predictive structure-based model (Challis *et al.* 2000) divides amino acids into groups by the chemical properties of their sidechains. First of all either polar or non-polar sidechains, then those subsets are further divided. The model depends on visualizing the binding pocket (Figure 3-4) and the effect of the residues at the identified critical positions. `DeepView` (Guex and Peitsch 1997) has been used in this work to visualize the binding pocket which harbours these residues. Residues were 'mutated' *in silico* to visualize the unknown binding pocket. In cases of doubt about the specificity of a binding pocket involved in a biosynthetic cluster which has been studied in detail, the view has been illustrated in a flattened form as in the structure-based predictive model (Challis *et al.* 2000).

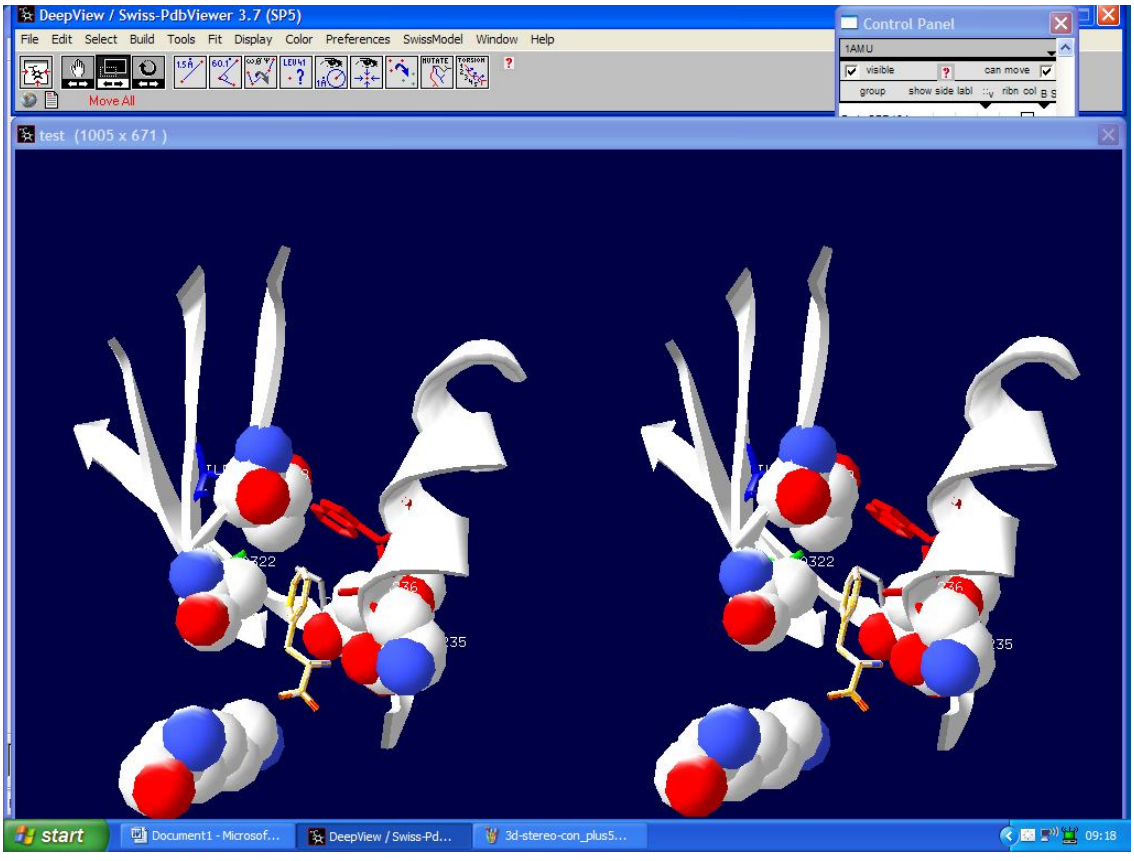


Figure 3-5 Working view of binding pocket from 1AMU visualized in DeepView. Critical residues thought to be involved in conferring specificity shown as space-filling models, backbone-only of residues scaffolding the binding pocket, bound ligand and AMP as stick views.

If the specificity seemed uncertain following categorization using the structure-based predictive model, additional techniques were used. Multiple sequence alignments were constructed including 1AMU, the unknown domain, and three sequences conferring each of two possible specificities to test whether either seems closer when aligned.

The protein sequence of the subsection of the unknown adenylation domain was used in a blastp search of INSDC to identify domains likely to be homologous. This is not ideal, as the specificity-conferring residues appear to undergo convergent evolution due to the functional constraint, hence those sequences closely related through evolution of the scaffold region do not necessarily share similar specificity. Such convergent evolution does not obviously affect higher level groupings, as amino acid binding domains appear to be monophyletic from the brief study described in **6.2.1Method development** and illustrated in **Figure 6-1**.

Adenylation of aryl acids

The crystal structure of the adenylation domain from 2,3-dihydroxybenzoic acid-AMP ligase DhbE from *Bacillus subtilis* (May *et al.* 2002) (PDB:1MD9) has been used in this work to refine alignments of adenylation domains likely to activate aryl acids. The residues identified by May and colleagues as having a role in aryl acid binding have been identified and the specificity of domains falling into this category has been assigned as they suggest, distinguishing between domains activating 2,3-dihydroxybenzoic acid and those activating salicylic acid (Figure 3-6).

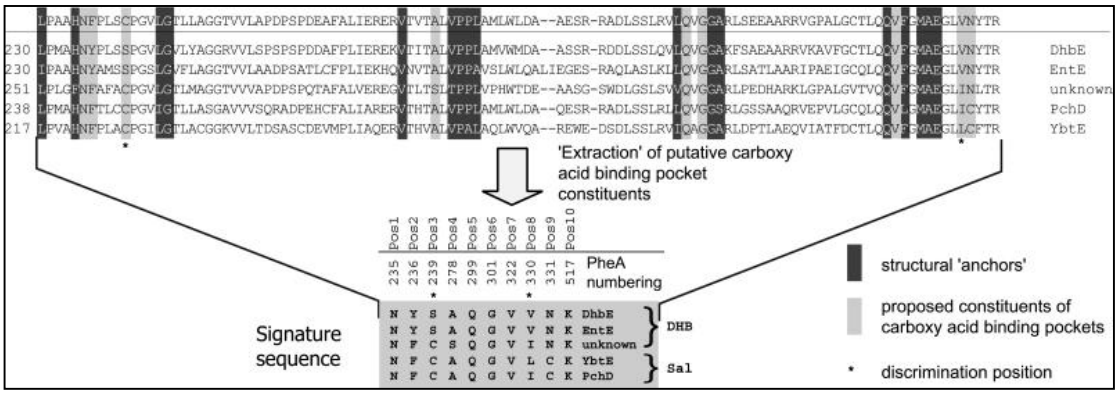


Figure 3-6 Signature sequence used to distinguish adenylation domains activating aryl acids. From May *et al.* 2002 GMAEG motif occurs at residues 331-335 of 1MD9; 2,3-dihydroxybenzoic acid (DHB)-activating domains differ consistently from those activating salicylic acid (Sal).

Adenylation domains activating other substrates

Alignments of adenylation domains representing various kinds of adenylation domains were used to form a tree and the presence or absence of conserved motifs identified in the literature decribed above was used to predict which substrate is likely to be bound by the protein domain, as described in **6.2.1Method development**.

3.2.2.2 Peptidyl carrier proteins

The excised PCP domain for which an NMR solution structure is available, tyrocidine synthetase TycC3 in *Brevibacillus brevis* (Weber *et al.* 2000), has been used at the model carrier protein domain (PDB: 1DNY) for alignments of this domain. Carrier domains have been shown to have an invariant serine (Lambalot *et al.* 1996) at positions aligning with residue 45 of this sequence (illustrated in Figure 3-8 below) which helps to identify this sequence type. Even low levels of sequence conservation appear to correctly predict function for this domain (Weber *et al.* 2000).

Sequences matching the Pfam phosphopantetheine attachment site model PF00550, and found adjacent to predicted proteins carrying domains characteristic of NRPS function, were extracted and aligned with the reference sequence 1DNY. Residues around the position aligning with Ser45 (PDB: 1DNY) in a multiple sequence alignment were examined. In this work, domains where the motif has the form **GxHSx** have been predicted to carry L-amino acids and domains where the motif is **GxDSx** predicted to carry the D-amino acid.

```
>1DNY:A|PDBID|CHAIN|SEQUENCE
   MPVTEAQYVA PTNAVESKLA EIWERVLGVS GIGILDNFFQ
   IGGHSLKAMA VAAQVHREYQ VELPLKVLFA QPTIKALAQY
   VATRSHHHHH H
```

Figure 3-7 Primary sequence of PCP domain 1DNY showing conserved Ser where phosphopathetheinylation occurs.

3.2.2.3 Condensation, Racemizing and Cyclizing domains

Domains matching PF00668

Structural data about NRPS condensation domains comes from the crystal structures from vibriobactin synthetase VibH from *Vibrio cholerae* (Keating *et al.* 2002) (PDB: 1L5A); a bidomain PCP-C structure from tyrocidine synthetase in *Brevibacillus brevis* which catalyzes heterocyclization (Samel *et al.* 2007) (PDB: 2JGP); and a four domain module (PDB: 2VSQ) from surfactin A synthetase of *Bacillus subtilis* (Tanovic *et al.* 2008).

Some researchers have identified a core motif for condensation as **HHxxxDG** (Stachelhaus *et al.* 1998), other researchers have found **HHxxxD** (Roongsawang *et al.* 2003). In heterocyclization (Z) domains **DxxxxD** may be found (Keating *et al.* 2002) and the involvement of several other residues has been demonstrated (Marshall *et al.* 2002). Dual function domains catalyzing both racemisation and condensation of the residue, are a subset of the donor D (dC) subfamily (Rausch *et al.* 2007) and may have motif **HHI/ LxxxxGD** at the equivalent position (de Bruijn *et al.* 2008). Domains matching the Pfam model PF00668 have thus been tested via a multiple sequence alignment to predict which condensation subfamily they fit into.

N-Methylation

Modules incorporating *N*-methylated residues have an additional domain between adenylation and carrier domains (Konz and Marahiel 1999), such as those in the enniatin (Haese *et al.* 1993; Pieper *et al.* 1995) and cyclosporin A (Billich and Zocher 1987) synthetase systems.

These domains are summarized by the Pfam A models PF08241 and PF08242. Motifs common to *N*-methyltransferase domains (Konz and Marahiel 1999) include ***M1*** VL(DE)GxGxG (assumed to be binding site for methyl donor S-adenosyl methionine (Konz and Marahiel 1999)); ***M2*** NELSxYRYxAV; ***M3*** VExSxARQxGxLD.

3.2.2.4 Thioesterase

Thioesterase domains matching Pfam model PF00975 (Schneider and Marahiel 1998) have been identified and the Ser/Asp/His catalytic triad checked by multiple sequence alignment using the methods of previous investigators (Samel *et al.* 2006).

3.2.3 Polyketide Synthase (PKS) domains

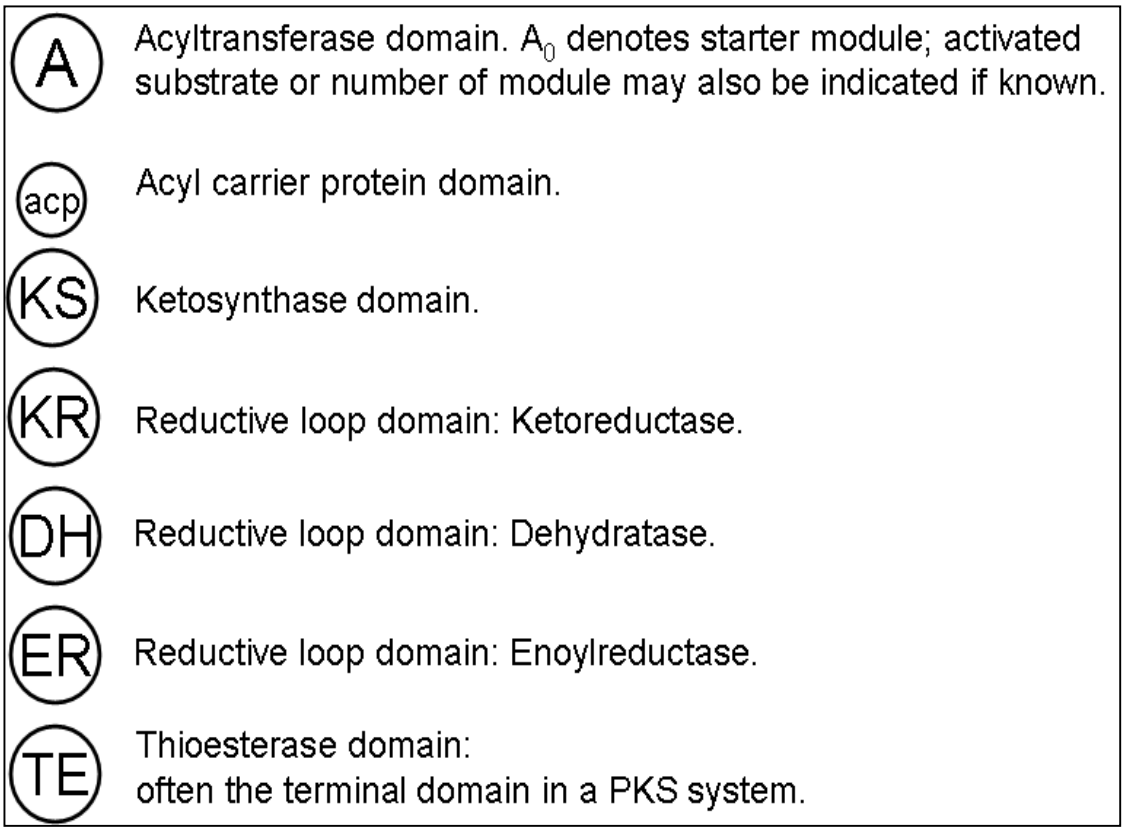


Figure 3-8 Key to polyketide synthase module maps.

3.2.3.1 Acyltransferase (AT) domains

Two crystal structures were used for aligning acyltransferase domains of PKS systems in this work. FabD from Escherichia coli (Serre *et al.* 1995) is the [acyl-carrier-protein] *S*-malonyltransferase (EC 2.3.1.39; PDB: 1MLA), and its orthologue malonyl-acetyltransferase (MAT) in *Streptomyces coelicolor* A3(2) (Keatinge-Clay,

A. T. *et al.* 2003) (PDB: 1NM2; SCO2387). PKS AT domains were identified at first by their match to Pfam model PF00698, and important residues are referred to in this work as numbered by alignment with 1MLA.

Several short sets of residues, apparently located in space at the active site of the acyltransferase domain, and thought to control the specificity of substrate in AT domains (Haydock *et al.* 1995; Lau *et al.* 1999) were check with multiple sequence alignments. A motif **HAF*H*/YAS*H*** is capable of affecting specificity between malonyl-coA and methylmalonyl-coA, where the underlined residues represent the active site His201 (Del Vecchio *et al.* 2003) and this was used to distinguish those two specificities. Ethylmalonyl-coA is selected by enzymes mostly having a small residue (Gly or Ala) at position 200 (Reeves *et al.* 2001) so this was used as the determining feature in predicting that specificity. Domains selecting methoxymalonyl-coA-derived extender units seem to have a conserved **hydrophobic-x-Trp** motif at positions aligning with FabD 188-190 (Haydock *et al.* 2005).

3.2.3.2 Ketosynthase domains

Conserved domain models from PFAM, PF00109 (N-terminal part) and PF02801 (C-terminal part), encompass the ketosynthase domain which catalyzes extension of the polyketide chain.

In the chain length factor characteristic of Type II PKS systems, a highly conserved glutamine replaces the KS active site cysteine (Bisang *et al.* 1999) so this was used as the determining factor in predicting Type II rather than Type I activity for a gene cluster. Mutations C169S, H309A, K341A, and H346A in the model Type II PKS system, actinorhodin, are blocked in early steps of the catalytic cycle (Dreier and Khosla 2000) which suggests their importance.

3.2.3.3 Reductive loop domains

Ketoreductase (KR)

Matches are found to PF00106, short chain dehydrogenase. Structure-based alignment can be performed with reference to actinorhodin KR (Korman *et al.* 2004) (SCO5086; PDB:1X7G) and to the crystallized KR domain of the first module of the

erythromycin synthase system (Keatinge-Clay, A. T. and Stroud 2006) (PDB: 2FR0). The catalytic triad KSY for KR modules predicted to be functional (Wu *et al.* 2005) were identified based on multiple sequence alignments.

Not all domains matching PFAM's conserved domain model PF00106 have the correct residues to predict KR activity, (for example: TmcA KR3 (Keatinge-Clay, A. T. and Stroud 2006)) so it is particularly necessary that the catalytic triad residues are checked in this domain. In some PKS modules there is no KR domain apparent; it has been assumed that in such circumstances there is no keto-reduction of the attached substrate, unless a freestanding keto-reductase is also found in the cluster. Where freestanding domains are found structural prediction has not been attempted due to the high number of possible products.

Stereospecificity of the alcohol after ketoreduction is controlled in the KR domain and domains have been predicted to have A or B function (*sensu* Caffrey 2003) by examining the characteristic residues (Caffrey 2003).

Dehydratase (DH)

DH domains can vary a great deal in the degree of sequence conservation observed (example: DH domains showed only 57-76% amino acid identity compared with each other in the tautomycin biosynthetic gene cluster (Choi, S. S. *et al.* 2007)). Important residues for DH domains were identified from the alignments published by previous researchers (Aparicio *et al.* 1996) but no Pfam A model appears to pick up this domain. The active site motif for this domain can be represented **HxxxGxxxxP** (Bevitt *et al.* 1992; Donadio *et al.* 1992), other investigators have added other conserved residues: **LxxHxxxGxxxxP** (Zirkle *et al.* 2004), and **HxxxGxxxxPG** (Wu *et al.* 2005). Inactive DH domains vary widely, but absence of the active site His residue is observed, for example in the inactive DH domains of the tautomycin biosynthetic cluster: **LxxPxxGxxxxP** in TmcA DH1, and **LxxYxxxGxxxxP** in TmcA DH2 (Choi, S. S. *et al.* 2007).

A search was conducted in `Artemis` for amino acid motifs varying in specificity between **HxxxxxxxxP** and **WLxxHxxxGxxxxPGxxxVxxxxxAxxxxG**. These searches for amino acid motifs were tested by trial and error against reference domains of known functionality from the rapamycin biosynthetic system (Schwecke

et al. 1995; Aparicio *et al.* 1996) to refine the search motif. To search for an amino acid string repeatedly in `Artemis: Goto > Navigator > Find amino acid string; Create Feature > misc_feature` with informative `/note="DH"`; search again with same string, create next feature, repeat until satisfactory set is obtained.

EnoylreductaseEnoylreductase (ER)

Amino acid strings from known ER-functional modules (rapamycin modules 1,3,7,13; concanamycin module 10) were aligned with regions of predicted proteins with possible ER function. The length of a PKS module alone may also be a clue to whether there are reductive loop domains present or not, but then the question of functionality still remains. ER domains were found to have 64-74% amino acids identical to each other within the tautomycin cluster (Choi, S. S. *et al.* 2007).

An invariant sequence **GGVGMAATQIA** was reported for ER domains within the rapamycin gene cluster (Aparicio *et al.* 1996), and a consensus **GGVGxAAxQxA** in the tautomycin gene cluster (Choi, S. S. *et al.* 2007). However, since these match the binding site for the NADPH coenzyme **GxGxxAxxxA** (Scrutton *et al.* 1990) they are not of unique predictive value and hence have not been used in this work.

3.2.3.4 Thioesterase

These domains were found using Pfam model PF00975. Several sequences are available from crystal structures of TE domains associated with polyketide synthases (Tsai *et al.* 2002) (PDB: 1NMA, 1MO2 and related sequences at various pH). Not all polyketide products are released by terminal thioesterification so the absence of such a domain does not indicate lackof function in a cluster.

3.3 Structure and pathway of likely products

3.3.1 The same, or different, product?

Products have been predicted to be the same where the same number of predicted domains, matching the same `Pfam` models, are present. In multienzyme proteins, products are predicted to be the same if the same predicted domains are present in the

same order as a characterised protein. If predicted domains are found in a cluster and not located on a multienzyme protein, it is assumed that the order of the coding sequences (and thus the order of enzyme domains) doesn't affect the biosynthetic product of the cluster.

Where enzyme domains known to be involved in biosynthesis in a related system are not found in an apparently related system in *S. scabies* 87.22, it has been assumed the biosynthesis cluster is non-functional or makes a different product.

Where possible, predictions of the chemical structure of the end product or pathway intermediates and an illustration of the biosynthetic pathway are provided. These have been illustrated using CambridgeSoft ChemDraw.

4 Results – genome overview

4.1 Introduction

This chapter contains an overview of findings from annotation of the complete genome sequence of *Streptomyces scabies* 87.22. These are placed in the context of other sequenced genomes. Comparisons are drawn primarily against the other two available streptomycete genomes, *Streptomyces coelicolor* A3(2) (Bentley *et al.* 2002) and *S. avermitilis* MA-4680 (Ikeda *et al.* 2003).

Results are also presented in this chapter from a study of the sequences likely to encode enzymes involved in complex product biosynthesis and which appear to be conserved across the three available streptomycete genomes. Gene clusters likely to encode enzymes for biosynthesis of complex natural products only found in *S. scabies* 87.22 of the three are described in Chapter 6.

4.2 Results

4.2.1 Method development

Links to relevant literature sources with evidence supporting functional annotation were added to the annotation file. This is an attempt to show the reasoning behind annotation decisions. The /citation qualifier was used even though this has a slightly different conventional usage. The PubMed **http://www.ncbi.nlm.nih.gov/pubmed/** accession number of relevant literature was appended to this qualifier. This data will be lost for the main public version of the genome because it is not part of the standard annotation submitted to the INSDC databases, but will be available upon request to the curators of the genome.

Stable RNA annotation

A new expert-curated database of tRNAs (Abe *et al.* 2009) will be a good resource for future annotation projects. More recent developments (Washietl and Hofacker 2004; Washietl *et al.* 2005) may be of interest for future annotations.

Coding sequence numbering

Missing entries in the numerical series will be apparent on inspection of the annotation: coding sequences predicted by the `Glimmer3` algorithm (Salzberg and Delcher 2004) and judged to be false positives (for method details see **2.3.2**). This method allows insertion of coding sequences judged by the annotator to be false negatives created by `Glimmer3`. These are created by referring to the FRAME plot (Bibb *et al.* 1984) and by overlay of `tblastx` against UNIPROT (Uniprot 2007) (See 2.5.3 Curation of CDS prediction). Such false negative marking also allows generous annotation of pseudogene fragments which would not be expected to completely conform to profiles of coding sequences due to degeneration having begun after relief of selective pressure, and which may be of use for comparative studies.

Numbering in tens may not be the most popular choice because it is not as straightforward as consecutive numbering, but it has the advantage of maintaining machine-sortable order including sequences annotated after the first pass. This convention is arguably a more accurate reflection of the open-ended nature of the process by which coding sequences are chosen by the annotator for inclusion in any particular release of the annotated genome. No data is lost as unedited versions of the automated coding sequence prediction are archived and extra sequences inserted are easily identified because the digits end in 2 or 3 instead of 1 as the original set do.

Coding sequences with identifier digits ending in 2 or 3 appear where `Glimmer3` has been judged by the annotator to have called a false negative, put in place by examination of the FRAME plot (Bibb *et al.* 1984) and by overlay of `tblastx` against UNIPROT (Uniprot 2007) (See 2.5.3 Curation of CDS prediction).

4.2.2 Genome overview

4.2.2.1 Genome size

The finished complete genome of *S. scabies* 87.22 consists of 10 148 695 base pairs. This is 1 481 188 base pairs larger than *"S. coelicolor"* A3(2) (Bentley *et al.* 2002) and 1 123 087 base pairs larger than *S. avermitilis* MA-4680 (Ikeda *et al.* 2003). The genome is linear and did not appear to include any extrachromosomal DNA (D. Harris and K. Seeger pers. comm.).

Estimates of the size of this genome before sequencing, for example by pulse field gel electrophoresis, may have been misleading because of the exceptionally high G+C content of genomes in this genus. Typically around 70% of bases are G+C, 92% of third 'wobble' position in codons (Nakamura *et al*. 1998; Nakamura *et al.* 2000). The genome of *S. scabies* 87.22 was estimated at 8.5 M base pairs (S. D. Bentley pers. comm.) before sequencing. Because G+C base pairs are an order of magnitude more thermodynamically stable than A+T pairs (Pranata and Jorgensen 1991), G+C pairs are likely to hold together more effectively under electrophoresis. The reduced friction from these stable G+C pairs - in comparison to unbiased DNA - may have allowed the whole genome to migrate more rapidly than would be expected, especially if the molecular mass markers used were of unbiased composition.

Streptomyces scabies 87.22 at approximately 10.1 M base pairs is the largest of the three *Streptomyces* complete genomes used in this work. All three are large in comparison to those of most micro-organisms (mean of finished microbial genomes sequenced by Wellcome Trust Sanger Institute to date = 3.9 M base pairs; "*S. coelicolor"* A3(2) ~8.7 M base pairs; *S. avermitilis* MA-4680 ~9.0 M base pairs). These large genomes are likely to have selective advantage, allowing the retention of morphological differentiation and capacity to sense and regulate growth in the widely fluctuating conditions of the soil niche. There is some evidence that deletion bias will remove non-advantageous DNA from a bacterial genome (Lawrence, J. G. and Hendrickson 2005), so it seems reasonable to assume that the large genomes are retained due to selective advantage rather than by lack of constraint.

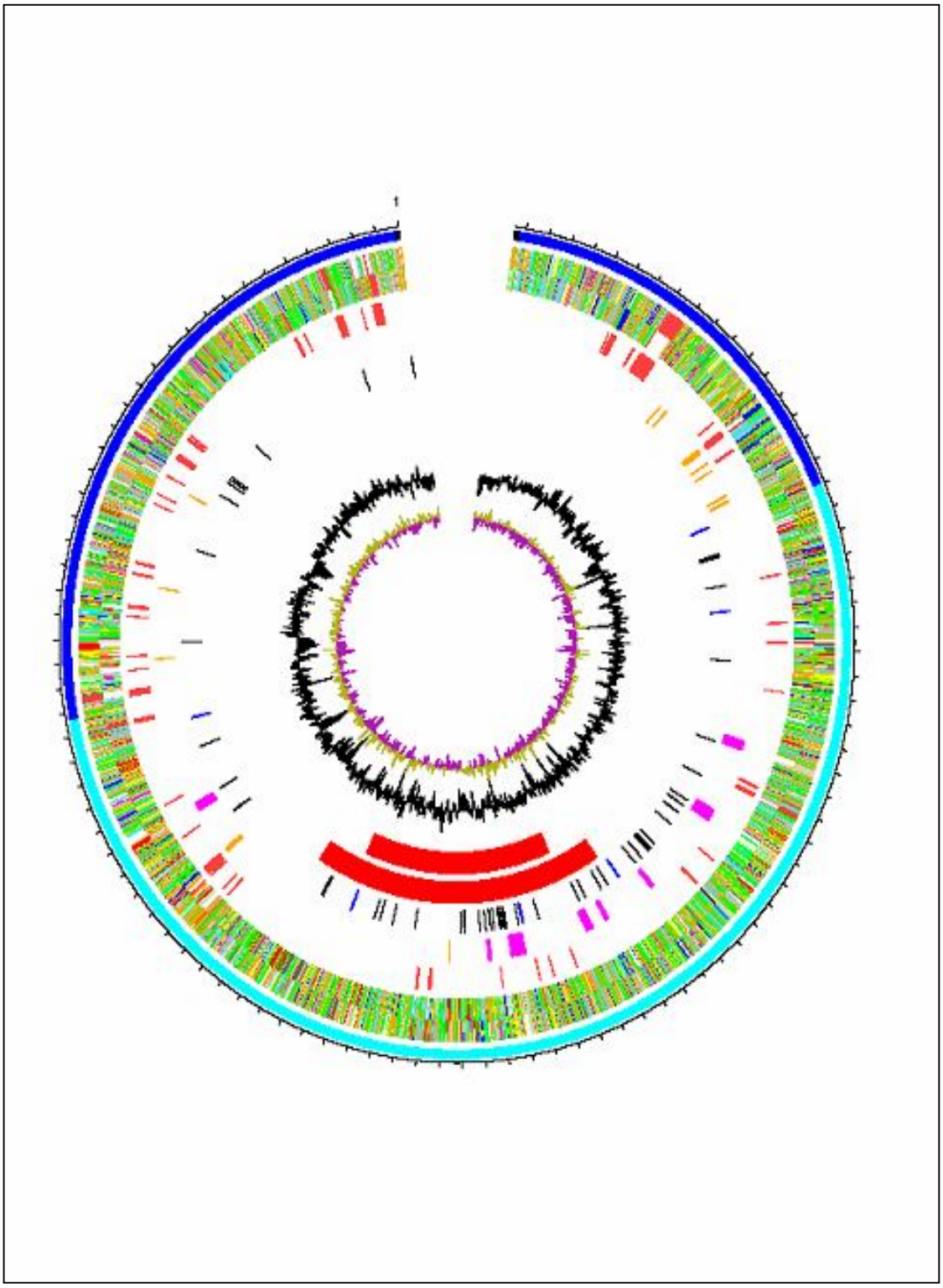


Figure 4-1 Overview of *S. scabies* 87.22 complete genome sequence. Circle 1: core (cyan) and arm (dark blue) regions of the chromosome marked with digits indicating Mbp. Circles 2 and 3, all CDS features with colour indicating functional annotation according to the scheme described in Methods **2.5.4** Circle 4, coding sequences thought to be involved in complex product biosynthesis. Circle 5, possible mobile element (magenta) and positions of pathogenicity features (gold). Circle 6, tRNA (navy) and rRNA (black) features. Circles 6 and 7, red bars indicate positions of inversions of *S. avermitilis* MA-4680 and "*S. coelicolor*" A3(2) genome synteny. Circle 8, GC content graph, circle 9, GC skew, magenta=positive, ochre=negative.

4.2.2.2 Coding sequences and probable pseudogenes

Coding sequences in the genome of *S. scabies* 87.22 are predicted to number approximately 8810, of which 63 are predicted as pseudogenes. These 63 are judged as unlikely to result in protein products because they lack any START codon, or include blocks to expression such as include frameshifts or inframe STOP codons (see further **Methods 2.5.3.6**).

This is not a much larger number of likely pseudogene sequences than were identified in the annotation of *"S. coelicolor"*A3(2) (Bentley *et al.* 2002) - see comparison Table 4-3. In some obligate pathogens genome reduction is apparent (Cole *et al.* 2001; Akama *et al.* 2009), and has been inferred if large numbers of pseudogenes are found. Since bacterial genomes appear to have a deletion bias (Mira *et al.* 2001) selecting against non-functional genetic material, the retention of pseudogenes is interpreted as very evolutionarily recent removal of selective pressure from those sequence, for example from the change in lifestyle due to obligate pathogenicity. The presence of relatively low numbers of pseudogenes in *S. scabies* 87.22, similar to the level at which they were found in non-pathogenic "*S. coelicolor"* A3(2).

4.2.2.3 No significant match

In *S. scabies* 87.22, 497 coding sequences were identified as having no significant degree of matching residues with any other predicted proteins so far submitted to UNIPROT (Apweiler *et al.* 2004; Uniprot 2007). These coding sequences could be of interest to future investigators because some will encode highly divergent proteins atypical of bacterial genomes, and it is possible that others may encode undiscovered gene families. For example, such sequences could be unknown because they encode enzymes with strain-specific functions that including pathogenicity. Similarity to known protein folds could be investigated using position specific iteration, (Altschul *et al.* 1997; Altschul and Koonin 1998), position hit iteration (Altun *et al.* 2006) and other techniques used for in-depth investigation of more divergent proteins, such as those in eukaryotic genomes.

4.2.2.4 Possible regulators

The proportions of coding sequences identified as having a regulatory function is similar, in "*S. coelicolor*" A3(2), and *S. scabies* 87.22. Approximately ten percent of the genome of *S. scabies* 87.22 is annotated as likely to encode regulatory sequences, with perhaps an additional two percent of coding sequences likely to encode a protein with a DNA binding domain, but not characterised as a regulator. This high number of putative regulators reflects the finding in "*S. coelicolor*" A3(2) which has been suggested to reflect *Streptomyces* soil niche adaptation (Bentley *et al.* 2002), since these organisms appear to have specialized in a great diversity of genetic material and fine control of the activity of these sequences.

It should be noted that since the larger genome of *S. scabies* 87.22 has proportionally similar numbers of putative regulators to that of "*S. coelicolor*" A3(2), there are many more sequences in this organism with this proposed function in actual numbers. Perhaps more regulatory proteins are required to regulate and co-ordinate the greater number of coding sequences in the genome of *S. scabies* 87.22. The genetic loci involved in pathogenicity will obviously have a regulatory component: additional regulatory functions will be required to sense the presence of host organisms, co-ordinate inputs from sensing functions, and orchestrate responses, and the selective importance of this niche can explain the maintenance of many more regulatory sequences.

4.2.2.5 Sub-cellular location

Secretion is of importance in the lifestyle of organisms in the *Streptomyces* genus known to have a saprophytic habit. Over 800 secreted proteins were predicted from the genome of the model organism, "*Streptomyces coelicolor*" A3(2) (Bentley *et al.* 2002). Coding sequence predictions in *S. scabies* 87.22 were classified for sub-cellular location as membrane proteins, secreted proteins, or (the majority) cytosolic by the methods described (**Methods 2.5.4.2**).

The proportion of coding sequences predicted to localise to the cell surface is similar in the two annotations, 26% of the genomic complement in *S. scabies* 87.22 and 30 % in "*S. coelicolor*" A3(2). A great deal of further investigation could be undertaken upon sequences identified as secreted, including those likely to be exported via twin

arginine translocation pathway, but such an investigation is outside the scope of this work.

Differences in prediction methods and interpretation during curation of the genome annotation could account for the small difference in numbers of apparently secreted proteins in *"S. coelicolor"* A3(2) and *S. scabies* 87.22; neither is based on experimental data from a survey of secreted proteins. All annotation is guesswork and inevitably the work of different curators will produce different results, so comparisons between the annotations can only indicate broad trends and obvious differences for further investigation. In order for valid comparisons to be made between genomes, identical methods would have to be used on both. A difference in the numbers and kinds of secreted proteins predicted - or confirmed if proteomics of the secreted fractions were obtained - would need to be specifically tested before any reliance could be placed on a difference between streptomycete genomes.

4.2.2.6 Relationship to other organisms

Ribosomal DNA comparison

S. scabies 87.22 has six copies of the operon encoding ribosomal RNA, as does *"S. coelicolor"* A3(2) (van Wezel *et al.* 1991) and *S. avermitilis* MA-4680. As in the other organisms, the six 16S rDNA copies in *S. scabies* 87.22 are not completely identical. Differences have been found in sequences preceding rDNA operons in *"S. coelicolor"* A3(2) (van Wezel *et al.* 1991). It may be that there are similar differences between copies within *S. scabies* 87.22 in the upstream region preceding 16S at the -600 to -250 positions.

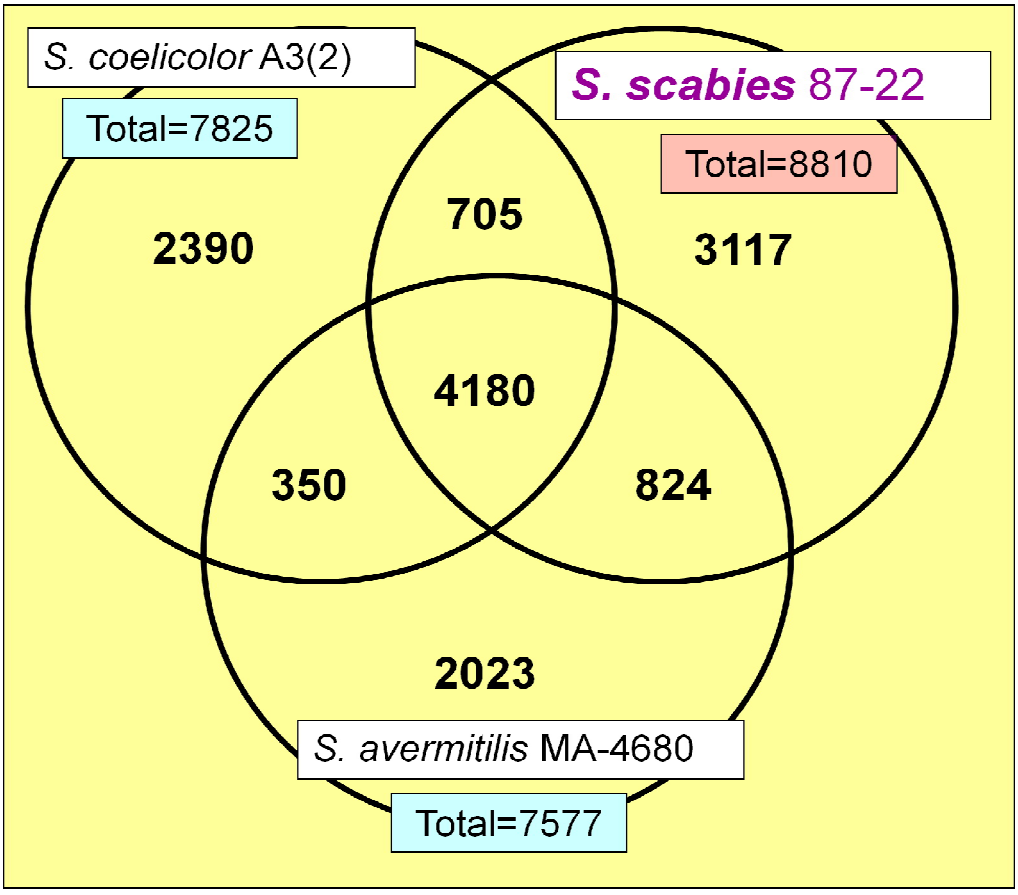


Figure 4-2 Venn diagram showing numbers of coding sequences apparently conserved across three streptomycete genomes by OrthoMcl results.

Orthomcl analysis of conservation

Orthomcl (Li, L. *et al.* 2003; Chen *et al.* 2006) was used to suggest likely orthologous relationships between coding sequence predictions for *Streptomyces scabies* 87.22 and those published for "*S. coelicolor*" A3(2), and *S. avermitilis* MA-4680 (Figure 4-2). This method suggests that approximately a third of the coding sequences in each of these three genomes are not found in either of the other two. (*S. avermitilis* MA-4680, 27%; "*S. coelicolor*" A3(2) 31%; *S. scabies* 87.22, 35%).

The slightly greater proportion of coding sequences in *S. scabies* 87.22 which are not conserved across the three genomes in — 35% compared with 31% in "*S. coelicolor*" A3(2) — seems to be in proportion to the larger genome rather than representing any greater non-conserved capacity (calculation of proportion follows).

$$\frac{8810\ CDS\ in\ S.scabies\ 87-22}{7825\ CDS\ in\ S.coelicolor\ A3(2)}$$

$$\times\ (31\%\ \text{S. coelicolor A3(2)sequences not conserved})$$

$$\approx 35\%\ \text{S. scabies 87} - 22\ \text{sequences not conserved}$$

It would not be surprising if a larger proportion of genetic material in *S. scabies* 87.22 was found not to be conserved in the other two organisms. The additional pathogenic capability of *S. scabies* 87.22 might have been found to be reflected in a larger proportion of genetic material not shared with the other non-pathogenic streptomycetes.

The finding that approximately a third of coding sequences predicted in *S. scabies* 87.22 are not conserved, and that this appears to be the case in both of the other two available genomes, may allow inferences about the amount of undiscovered genetic material in the *Streptomyces* genus. Given the great importance of this genus as a source of producers of biologically active compounds (Berdy 2005), it is of interest to know how much more genetic material might be expected from members of this genus for which the complete genome sequence is not yet determined.

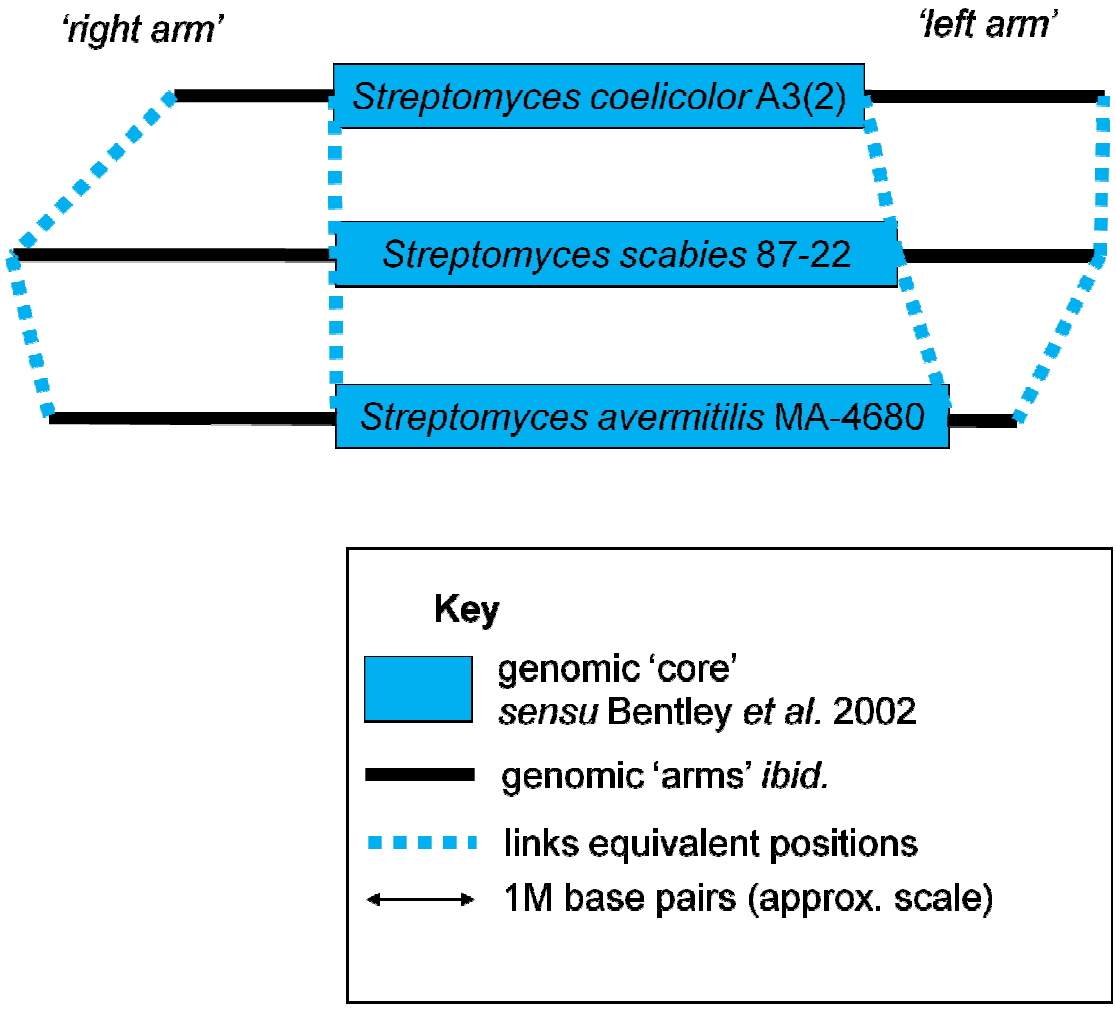


Figure 4-3 Core and arms layout of three streptomycete genomes.

4.2.2.7 Genome organization

Core/arms

Bentley and colleagues defined the core region in *"S. coelicolor"* A3(2) as the region lying between approximately the 1.5 Mth and 6.4 Mth base pairs, by reference to high-scoring pairs in alignment with the genome of *Mycobacterium tuberculosis*, the other actinomycete for which a complete genome had been sequenced (Bentley *et al.* 2002). Ikeda and colleagues identified a similar core region of the *S. avermitilis* MA-4680 genome by comparison to *"S. coelicolor"* A3(2). The core and arms in *S. scabies* 87.22 have therefore been defined for the purposes of this investigation by comparing regions of synteny between *"S. coelicolor"* A3(2), *S. avermitilis* MA-4680, and *S. scabies* 87.22. Positions of the core and arms in *S. scabies* 87.22 are presented in Table 4-1 and summarized in Figure 4-3.

organism	total size (/base pairs)	start core	end core	core size	left arm	right arm	arms total	% core	%arms
"S. coelicolor" A3(2)	8667507	1500000	6400000	4900000	2267507	1500000	3767507	56.5	43.5
S. avermitilis MA-4680	9025608	8404923	2734171	5670752	2734171	620685	3354856	62.8	37.2
S. scabies 87-22	10148695	8286903	2941783	5345120	1861792	2941783	4803575	52.7	47.3

Table 4-1 Comparison of positions for core and arms of three streptomycete genomes.

organism compared	Change in core size (/base pairs)	% of total increase in core	Change in size of arms (/base pairs)			% of total increase in arms
			left	right	total	
"S. coelicolor" A3(2)	445120	30.1	-405715	1441783	1036068	69.9
S. avermitilis MA-4680	-325632	-29.0	1241107	207612	1448719	129.0

Table 4-2 Calculations for position of additional material in S. scabies 87.22.

Analogous positions in *S. scabies* 87.22 for the core have been defined, by the extent of `blast` matches indicating conserved gene arrangement between the three genomes, such that the core is between base pairs 2941783 and 8286903. The core of the genome (*sensu* Bentley *et al*. 2002) is the region across which comparisons are useful between the three full genomes available, because there is broad synteny. Beyond the core region there is no apparent conservation of gene order. It may be that a high rate of loss and acquisition of genetic material occurs in these regions due to rearrangements such as arm exchange (Hopwood 2006) and thus it is difficult to make meaningful comparisons of the arm regions between strains.

Compared to "*S. coelicolor*" A3(2), there are around 445 k base pairs more in the core region of *S. scabies* 87.22. In *S. avermitilis* MA-4680, there appears to be a reduction of over 325 k base pairs in the core region. This means that in comparison between "*S. coelicolor*" A3(2) and *S. scabies* 87.22, just over 30% of the total increase falls within the core region.

Nearly 70% of the additional material in *S. scabies* 87.22 compared to "*S. coelicolor*" A3(2) is found in the arms regions at either side of the core of the chromosome. The left arm of *S. scabies* 87.22 is smaller than that of "*S. coelicolor*" A3(2), so the right arm of *S. scabies* 87.22 is the region having the greatest increase in quantity of DNA in comparison between the two genomes. In comparison with *S. avermitilis* MA-6480, both arms have increased in size in *S. scabies* 87.22, which has more than 1.2 M base pairs of additional material in the left arm and 207 k base pairs in the right arm. It is not possible with the available data to predict the order or direction of insertions or deletions, but further study of the locations of pathogenicity-related genetic material

might indicate insertions in the *scabies* lineage since pathogenicity is not common in the genus and would seem unlikely to be a conserved ancestral trait.

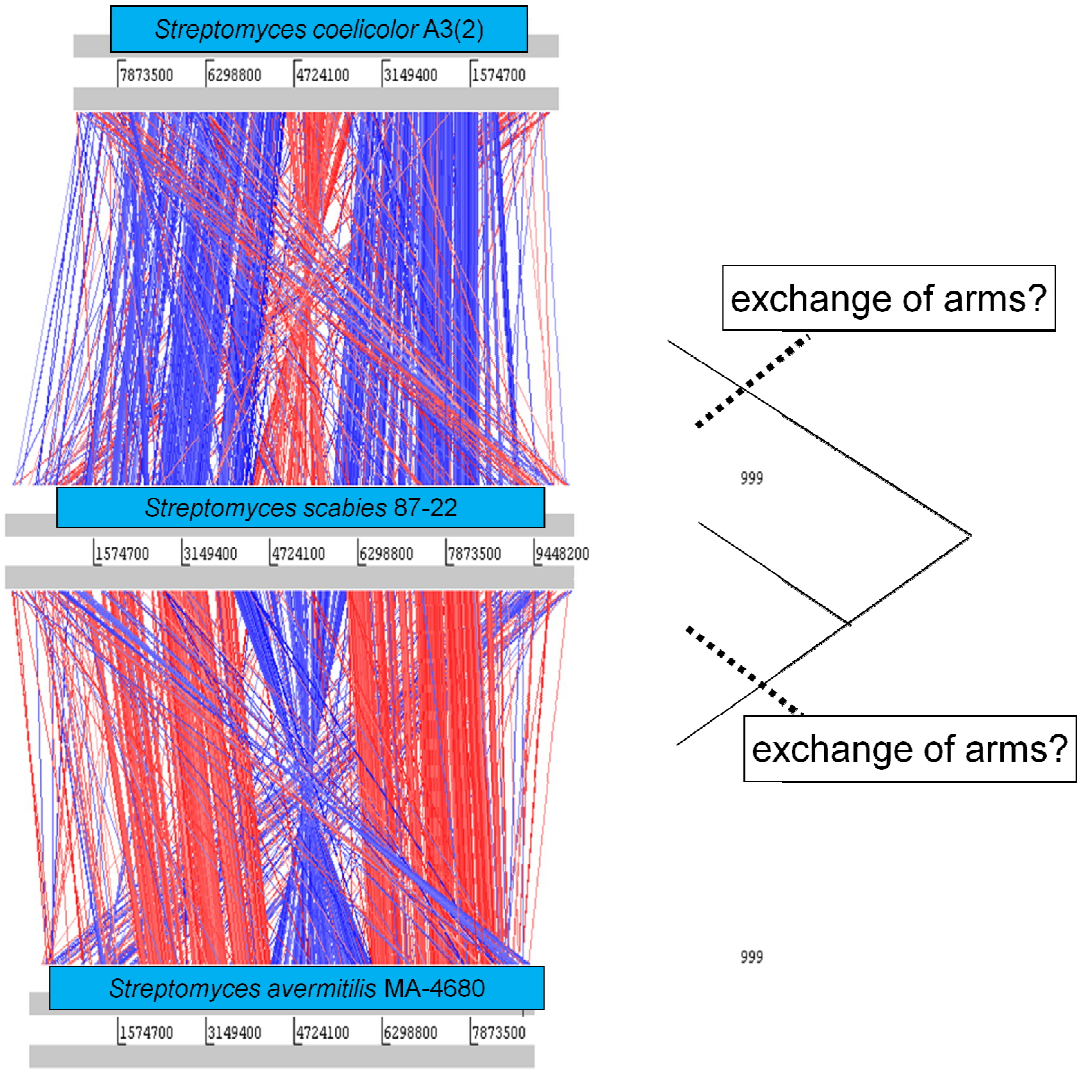


Figure 4-4 ACT comparison of *"S. coelicolor"* A3(2), *S. scabies* 87.22, *S. avermitilis* MA-4680 showing central inversion. Direct comparisons illustrated with red connecting lines between the sequences where alignments are found, and alignments reversed in sense are illustrated with blue connecting lines. The top comparison *("S. coelicolor"* A3(2) to *S. scabies* 87.22) has opposite colours to the lower comparison because it been flipped in orientation to show the regions of greatest synteny. Dendrogram of three streptomycete genomes: dashed lines indicate when rearrangement events may have resulted in inversion of the central region of the chromosome.

Central inversions

From comparison with blast visualised in ACT (Figure 4-4) it appears that the central regions of both *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680 are inverted compared to the *S. scabies* 87.22 sequence. Inverted regions are illustrated in blue,

direct matches in red. The two inversions – one in "*S. coelicolor*" A3(2) vs *S. scabies* 87.22, and one in *S. avermitilis* MA-4680 vs *S. scabies* 87.22, are in different places so it is inferred these must have occurred in different events. The simplest explanation is that S. *scabies* 87.22 retains the ancestral arrangement (R. Loria pers. comm.), with both of the other two genomes having survived the inversion of the central region of the chromosome. The inversion may have resulted from an exchange of replication forks, each of which proceeds along the linear genome during chromosome replication (Hopwood 2006).

4.2.2.8 Key features comparison

Terminal inverted repeats

The terminal inverted repeats (TIRs) at each end of the *S. scabies* 87.22 chromosome appear to appear to contain 15 coding sequences (

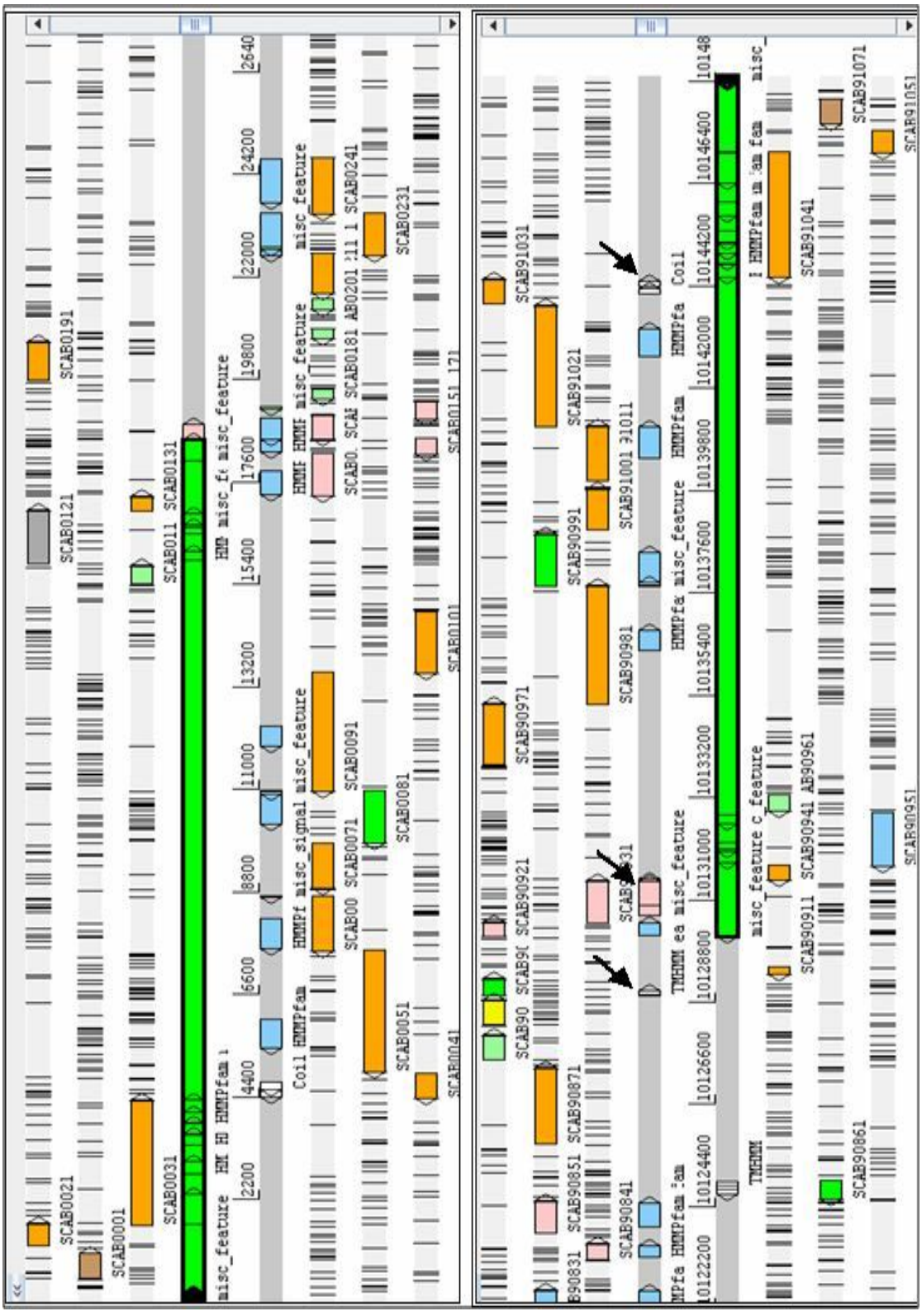


Figure 4-5). These inverted repeats are 18 486 base pairs long, hence are of middling length amongst those chromosome ends so far sequenced in the genus. The TIRs of *S. ambofaciens* DSM40697 are more than ten times larger than those of *S. scabies* 87.22 at around 210 000 base pairs and contain over 200 predicted coding sequences (Leblond *et al.* 1996). The TIRs of "*S. coelicolor*" A3(2) are similar to those of *S. scabies* 87.22 - just over three hundred base pairs larger. The TIRs of *S. avermitilis* MA-4680 were found to be very short (167 base pairs). The comparatively greater length of TIRs in the *S. scabies* 87.22 genome supports the observation that these short TIRs in *S. avermitilis* MA-4680 may be unusual in the genus. TIRs have been shown to vary between strains within a species (Choulet *et al.* 2006b) so conclusions cannot be drawn in general about these features between species.

Table 4-3 Key feature comparison of three streptomycete complete genomes. bp= base pairs

	"Streptomyces coelicolor" A3(2) (version published)	*Streptomyces avermitilis* MA-4680 (version 2007-05-07)	*Streptomyces scabies* 87-22
Whole genome size	8667507 bp	9025608 bp	10148695 bp
Structure	linear	linear	linear
Terminal inverted repeat (TIR) length	21653 bp	167 bp	18486 bp
G+C content	72.12%	70.70%	71.76%
Proportion of genome predicted coding	88. 90 %	86.20%	86.50%
Coding sequences (CDS) predicted	7825	7580	8804
... of which pseudogenes predicted	55	not published?	43?
Mean length of coding sequences	991 bp	1034 bp	1001 bp
Ribosomal RNA operons (16S-23S-5S)	6	6	6
Transfer RNAs predicted	63	68	75
Source	Bentley *et al.* , Nature, 2002 Genome project website: http://www.sanger.ac.uk/Projects/S_coelicolor	Ikeda *et al.* , Nature biotechnology, 2004 Genome project website: http://avermitilis.ls.kitasato-u.ac.jp/	Yaxley, Bentley, and Loria unpublished 2009 Genome project website: http://www.sanger.ac.uk/Projects/S_scabies

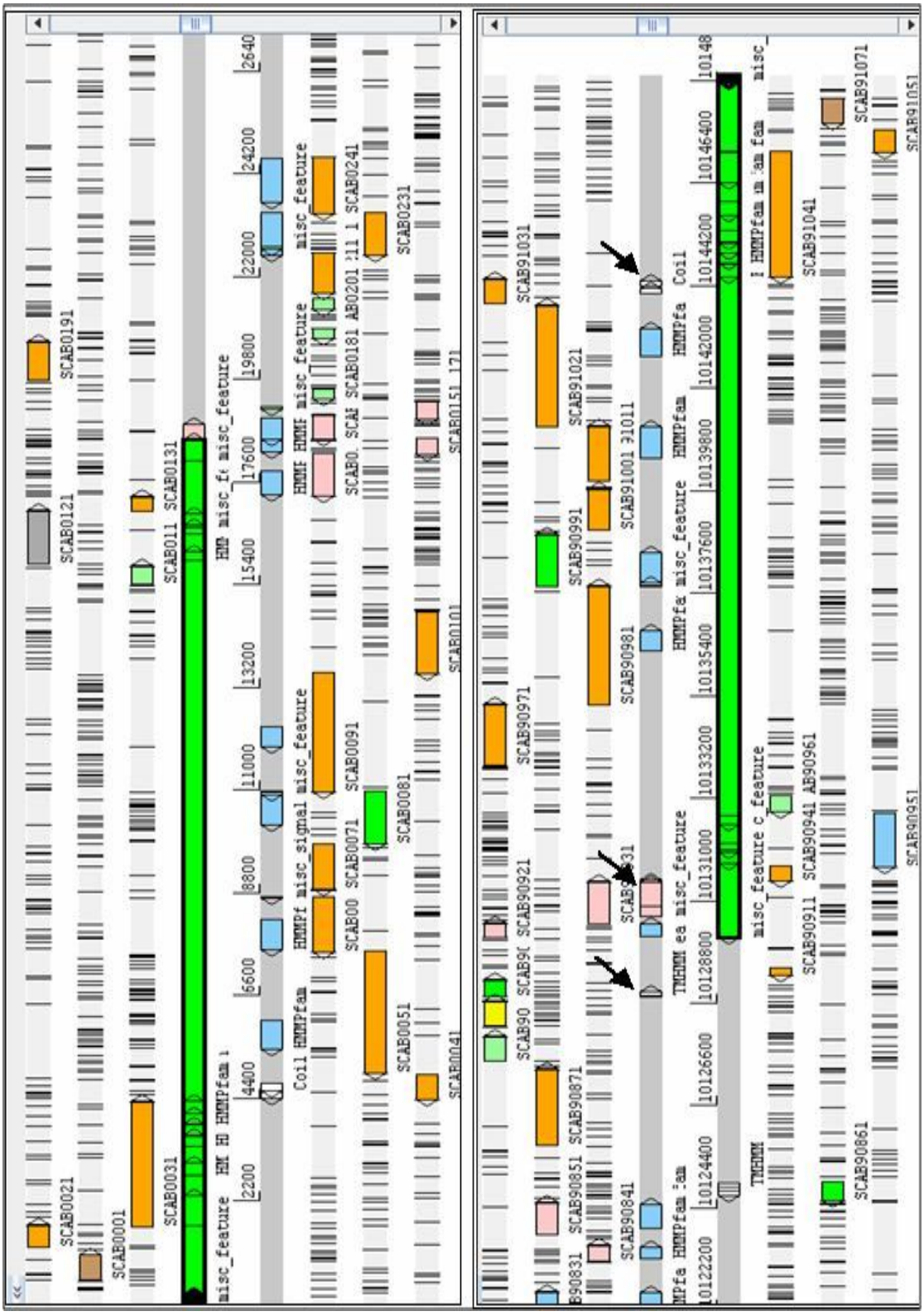


Figure 4-5 Terminal inverted repeats of S. scabies 87.22 visualized in Artemis. Top, left TIR, below, right TIR. The green feature with the bold black outline in the centre of the screen is the exact duplication of nucleotide sequence. Forward and reverse protein coding frames and the coding sequences predicted in them appear towards top and bottom of each screenshot, outwards from the nucleotide sequence along the centre of the screen. Features indicating conserved domains, transmembrane helices and other regions of interest in predicted coding sequences have been mapped along the central region (examples indicated by arrows).

Numbers of tRNAs

There are several more tRNA coding sequences in *S. scabies* 87.22 than in *"S. coelicolor"* A3(2) or *S. avermitilis* MA-4680. This appears to be in proportion (see calculation below) to the larger genome of *S. scabies* 87.22.

"S. coelicolor" A3(2) $\frac{8.67\times10^{6}\text{ base pairs}}{63\text{ tRNAs}} \approx$137 579 base pairs per tRNA

S. avermitilis MA-4680 $\frac{9.03\times10^{6}\text{ base pairs}}{68\text{ tRNAs}} \approx$132 730 base pairs per tRNA

S. scabies 87.22 $\frac{10.15\times10^{6}\text{ base pairs}}{75\text{ tRNAs}} \approx$135 200 base pairs per tRNA

The ratio at which tRNAs are found in the three genomes appears to be approximately the same, around one tRNA encoded per 135 000 base pairs. There might be some reason genes for tRNAs are maintained at this frequency, perhaps related to the observation that prophage insertion sites often lie within tRNA genes (Campbell 2003) and it is likely to be detrimental to the host bacteria to carry too great a cargo of DNA without selective advantage. Further studies could test whether the distribution of tRNA genes along the chromosome is random in general in bacterial genomes, which might give further clues about whether such a mechanism is active.

4.2.2.9 Other

BldA regulon

The rare leucine codon TTA has been implicated in developmental regulation in "*Streptomyces coelicolor"* (Takano *et al.* 2003; Hesketh *et al.* 2007; Li, W. C. *et al.* 2007). A computational script adapted from the one kindly provided by Li and colleagues was used to identify 248 TTA codons in the genome of *S. scabies* 87.22. The full list is appended (Appendix E, attached on digital medium) Some TTA locations are associated with gene clusters involved in complex product metabolism and pathogenicity (Table 4-4). A *bldA*-like phenomenon in *S. scabies* 87.22 has not yet been demonstrated, and as such it seems premature to speculate on the possible effects of TTA codons at these positions, but it would be interesting to find out whether there is an effect, given the location of these codons in the biosynthetic clusters for the key phytotoxins (thaxtomins) and in clusters for something like

pyochelin, the possible coronafacic acid-like product and the concanamycin cluster SCAB83871-SCAB84091.

TTA-containing CDS	#tta	nearby cluster CDS	product of cluster
SCAB1391	1	SCAB1411-SCAB1571	pyochelin-like?
SCAB1661	1	SCAB1411-SCAB1571	pyochelin-like?
SCAB3351	1	SCAB3281-SCAB3361	unknown
SCAB31801	1	SCAB31761-SCAB31841	thaxtomins
SCAB31971	1	SCAB31961-SCAB32051	lantibiotic?
SCAB32041	1	SCAB31961-SCAB32051	lantibiotic?
SCAB43931	1	SCAB43961	unknown
SCAB63131	1	SCAB63251-SCAB63401	unknown
SCAB69851	1	SCAB69771-SCAB69871	unknown
SCAB69871	1	SCAB69771-SCAB69871	unknown
SCAB79591	1	SCAB79591-SCAB79721	cfa-derivative?
SCAB79621	1	SCAB79591-SCAB79721	cfa-derivative?
SCAB84101	1	SCAB83871-SCAB84091	concanamycins

Table 4-4 TTA codons nearby or within gene clusters expected to direct complex product biosynthesis. First column: systematic identifier of coding sequence found to contain TTA; second column: number of TTA codons found within putative coding sequence; third column indicates estimated boundaries of the closest complex product gene cluster, fourth column indicates proposed product of gene cluster where known.

Streptomyces regulatory proteins

New protein families identified as a result of the "*S. coelicolor*" A3(2) complete genome sequencing project include a family with a conserved domain called "bacterial transcriptional activator" (Yeats *et al.* 2003). This protein family include known proteins such as the pleiotropic regulator of antibiotic production in "*S. coelicolor*" A3(2), AfsR (Horinouchi *et al.* 1990), as well as pathway-specific regulators in the both the actinorhodin (act) and undecylprodigiosin (red) biosynthetic pathways in "*S. coelicolor*" (Floriano and Bibb 1996).

The conserved architecture of this kind of protein consists of PF00486 towards the N-terminal of the primary sequence, which is helix-turn-helix clan domain, DNA-binding fold, with the newly discovered conserved domain PF03704 towards the C-terminal. There are eleven proteins in the *S. scabies* 87.22 genome with the architecture described (Table 4-5). These vary in size from 836 (SCAB63171) to 3380 base pairs (SCAB66351). Several appear to be associated with complex product biosynthesis clusters, others are not obviously in such positions and their function

remains to be discovered. SCAB20511 and SCAB20521 are part of a interesting cluster which looks as if it may have arisen from a tandeom duplication, and the tetratricopeptide repeat domains encoded there indicate possible function in protein association in a complex. SCAB40101 appears to be part of a sensor/regulator pair: further investigation might be merited as it is possible the regulator could activate multiple biosynthetic pathways as AfsR does in "*S. coelicolor*".

CDS	start	end	product of cluster
SCAB1371	142512	144437	pyochelin-like?
SCAB20511	2322714	2325803	no obvious cluster
SCAB20521	2325835	2327868	no obvious cluster
SCAB40101	4501260	4504229	two component sensor/regulator
SCAB51941	5776270	5779254	no obvious cluster
SCAB58571	6513419	6516964	inside transport system?
SCAB61801	6853696	6856770	no obvious cluster
SCAB63171	7000688	7001524	unknown product
SCAB66351	7341720	7345100	transport system?
SCAB84101	9396902	9398950	concanamycins

Table 4-5 Coding sequences in *S. scabies* 87.22 genome with significant scores to 'bacterial transcriptional activator' domain PF03704. Second and third columns, base position of coding sequence. Fourth column: function of gene cluster associated with the coding sequence, where known.

4.2.3 Complex product gene clusters

In this work 240 coding sequences have been identified as possibly encoding enzymes involved in biosynthesis of complex natural products. This is approximately 2.5% of the genome, one in forty of the coding sequences identified in this work a proportion that appears to be typical for the streptomycete genomes so far sequenced.

4.2.3.1 Conserved clusters

Of the gene clusters found in the complete genome sequence of *S. scabies* 87.22 and identified in this work as likely to encode enzymes involved in biosynthesis of complex natural products, ten appear to be conserved across the three genomes (**Error! Reference source not found.**), by methods described in **Chapter 3**. A further five clusters were found to be partly conserved, present in either "*S. coelicolor*" A3(2) or *S. avermitilis* MA-4680 as well as *S. scabies* 87.22, but not both (**Error! Reference source not found.**).

The two gene clusters found in *S. scabies* 87.22 and *S. avermitilis* MA-4680, but not in *"S. coelicolor"* A3(2) are: cryptic system including a multienzyme protein SCAB72991 with transmembrane domains, an adenylation-family domain, and a phosphopantetheine attachment site for carrier protein function; and a NIS (Challis 2005) biosynthetic system which may encode a siderophore.

Three clusters were found in *S. scabies* 87.22 which appear to be present in *"S. coelicolor"* A3(2) by similarity, but not in *S. avermitilis* MA-4680. *S. scabies* 87.22 has a type III PKS synthase gene for biosynthesis of germicidins. Biosynthesis of spore germination inhibitor germicidins was predicted from the presence of the biosynthetic gene and has been confirmed by LC-MS/MS and TOF (L. Song and G. L. Challis, pers. comm., data not shown).

In addition *S. scabies* 87.22 appears to share with *"S. coelicolor"* A3(2) a gene cluster likely to encode enzymes for biosynthesis of 2-methylisoborneol, SCAB5031-SCAB5051, with no obvious differences to clusters identified as functional for production of these compounds (Komatsu *et al.* 2008).

The gene cluster encoding enzymes for biosynthesis of exopolysaccharide EPS 139A characterised in *Streptomyces sp.* 139 and found in *"S. coelicolor"* A3(2) (Wang, L.-y. *et al.* 2003a; Wang, L.-y. *et al.* 2003b) is also present in *S. scabies* 87.22, but again missing from *S. avermitilis* MA-4680. These coding sequences in *S. scabies* 87.22 were annotated by comparison with the coding sequences submitted by Wang and colleagues (AY131229).

Clusters present in all three available genomes

Clusters which are found in all three of the available genomes could be conserved across the *Streptomyces* genus, because the three organisms – *"S. coelicolor"* A3(2), *S. avermitilis* MA-4680 and *S. scabies* 87.22 and might be found to be essential for the free-living streptomycete niche in soil.

Biosynthetic system	Metabolite	*S. scabies* systematic id	Size estimate /kb (0 d.p)	Conserved in "*S. coelicolor*" A3(2) and *S. avermitilis* MA-4680?
Phytoene/polyprenyl synthase	carotenoid pigment	SCAB5431-SCAB5511	12	both
Hopene/squalene synthase	**pentacyclic hopanoids**	**SCAB12951-SCAB13061**	**14**	**both**
NRPS-independent synthetase	siderophore?	SCAB18371-SCAB18421	7	both
Terpene synthase	geosmin	SCAB20121	2	both
NRPS-independent synthetase	siderophore?	SCAB24661-SCAB24751	16	both
Bacterial type II polyketide synthase	spore pigment	SCAB43271-SCAB43341	8	both
NRPS-independent synthetase	**desferrioxamines**	**SCAB57921-SCAB57981**	**9**	**both**
Tyrosinase	melanin pigment	SCAB59231-SCAB59241	2	both
Ectoine synthase	ectoine compatible solute	SCAB70711-SCAB70751	4	both
Tyrosinase	melanin pigment	SCAB85681-SCAB85691	2	both
Monoterpene cyclase	2-methylisoborneol	SCAB5031-SCAB5051	4	in "*S. coelicolor*" A3(2)
Polysaccharide	exopolysaccharide	SCAB23341-SCAB23551	29	in "*S. coelicolor*" A3(2)
Mixed	unknown complex product	SCAB72991	4	in *S. avermitilis MA-4680*
Bacterial type III polyketide synthase	**germicidins**	**SCAB80171**	**4**	**in "*S. coelicolor*" A3(2)**
NRPS-independent synthetase	siderophore?	SCAB84501-SCAB84521	8	in *S. avermitilis MA-4680*
Nonribosomal peptide synthetase	**pyochelin siderophore?**	**SCAB1411-SCAB1571**	**35**	**no**
Nonribosomal peptide synthetase	**thaxtomins**	**SCAB31761-SCAB31841**	**19**	**no**
Bacterial type I polyketide synthase	**concanamycins**	**SCAB83871-SCAB84091**	**95**	**no**
Nonribosomal peptide synthetase	lipopeptide?	SCAB3281-SCAB3361	34	no
Class II DAHP synthase	2-amino-3-hydroxybenzoic acid?	SCAB12021-SCAB12111	12	no
Nonribosomal peptide synthetase	unknown complex product	SCAB19681-SCAB19731	8	no
Lantibiotic	unknown lantibiotic	SCAB31961-SCAB32051	14	no
Hybrid NRPS/PKS system	unknown	SCAB43961	13	no
Hybrid NRPS/PKS system	unknown	SCAB62901-SCAB63011	14	no
Mixed	unknown complex product	SCAB63251-SCAB63401	8	no
Mixed	unknown complex product	SCAB69771-SCAB69871	13	no
Bacterial type I polyketide synthase	coronafacic acid derivative?	SCAB79591-SCAB79721	26	no
Hybrid NRPS/PKS system	unknown	SCAB78961-SCAB78981	8	no
Nonribosomal peptide synthetase	NRPS siderophore	SCAB85461-SCAB85521	42	no

Table 4-6 Gene clusters identified as possibly encoding enzymes for biosynthesis of complex natural products. Bold entries have evidence for production. Yellow-shaded entries (top ten) seem to be conserved across the three available genomes; the bottom fourteen appear to be only in S. scabies of the three and are described in greater depth in Chapter 6.

Gene clusters with known products judged to be conserved across the three organisms (*S. scabies* 87.22, *"S. coelicolor"* A3(2), *S. avermitilis* MA-4680) include a hopene/squalene synthase system SCAB12951-SCAB13061 likely to encode enzymes for biosynthesis of membrane-stabilising hopanoids (Poralla *et al.* 2000). SCAB20121 identified as *geoA* geosmin biosynthesis gene by similarity. The biosynthetic products from the gene cluster SCAB5431-SCAB5511 are expected to be light-induced and include isorenieratine and beta-carotene by similarity of the gene cluster to those found in *"S. coelicolor"* A3(2), *S. griseus* (Lee *et al.* 2001) and linear plasmid pSLA2-L of *S. rochei* (Mochizuki *et al.* 2003) (**Error! Reference source not found.**).

A PKS gene cluster for biosynthesis of spore pigments first identified in *"S. coelicolor"* (Davis and Chater 1990; Kelemen *et al.* 1998) appears to be common to all three streptomycete genomes available. It was not determined whether a possible operator region upstream of the -35 site in this cluster (Novakova *et al.* 2004) is conserved in *S. scabies* 87.22.

Melanins

Two *melC2C1* gene clusters are found in *S. scabies* 87.22 (**Error! Reference source not found.**). *S. scabies* has been found to produce melanins (Beausejour and Beaulieu 2004), like *S. avermitilis* MA-4680 (Omura *et al.* 2001), but unlike *"S. coelicolor"* A3(2), though genes for production of melanins are found in *"S. coelicolor"* A3(2) (Bentley et al. 2002). Tyrosinases involved in production of two kinds of melanins (Omura *et al.* 2001) in streptomycetes are of wider interest to biochemists as they have been used as models for the catalytic 'Type 3 copper centre' (Claus and Decker 2006). Structural models for the tyrosinase/caddie complex have been published (Matoba *et al.* 2006). In-depth comparison of the *melC2C1* clusters of *S. scabies* 87.22 and similar sequences might reveal details of the differences between functional and non-functional genes.

IucA/IucC family: possible synthetase enzymes

It was predicted from the genome sequence that *S. scabies* 87.22 encoded biosynthetic enzymes for desferrioxamines. An insertion sequence was found in this cluster (Figure 4-6) in comparison with the *des* cluster characterised in *"S.*

coelicolor" A3(2) (Barona-Gomez *et al.* 2004). This insertion sequence is not present at this position in the genome of *S. avermitilis* MA-4680 either, from my comparisons. Production of these compounds has been confirmed by LC-MS/MS and TOF (L. Song and G. L. Challis, pers. comm., Figure 4-7), and hence does not seem to be affected by the insertion.

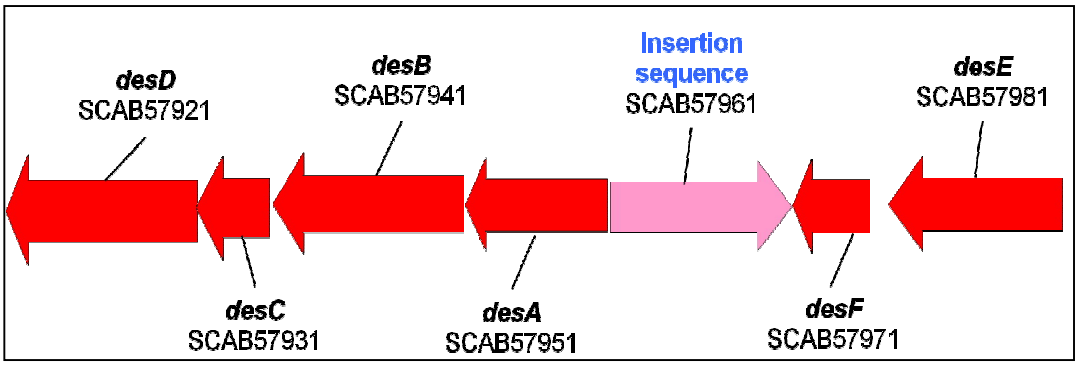


Figure 4-6 Summary of gene cluster for biosynthesis of desferrioxamines.

Desferrioxamines are important siderophores in clinical use for transfusion iron overload (Norman 1964). Production and utilization of desferrioxamines is found in many streptomycetes (Yamanaka *et al.* 2005) as well as other bacteria (Essen *et al.* 2007; Kadi *et al.* 2008; Zawadzka *et al.* 2009). Trimerization and macrocyclization of desferrioxamine E is directed by DesD (Kadi *et al.* 2007), a protein which carries the conserved domain PF04183 (IucA/IucC family (de Lorenzo and Neilands 1986)).

Several other gene clusters in the genome of *S. scabies* 87.22 (beside that which is predicted to direct biosynthesis of desferrioxamines) also carry the PF04183 conserved domain and possibly encode capacity for assembly of complex natural products. Three clusters containing PF04183 domains in the genome of *S. scabies* 87.22 are conserved across the three complete genome sequences available (SCAB18371-SCAB18421, SCAB24661-SCAB24751, and the desferioxamine cluster; **Error! Reference source not found.**). The gene cluster containing PF04183 is in both *S. scabies* 87.22 and *S. avermitilis* MA-4680 (SCAB84501-SCAB84521; **Error! Reference source not found.**). Detailed studies of gene clusters likely to encode complex product biosynthesis in this work have been limited to those clusters only found in *S. scabies* 87.22 of the three available complete streptomycete genomes. An in-depth study of these IucA/IucC-like clusters has not been undertaken in this work.

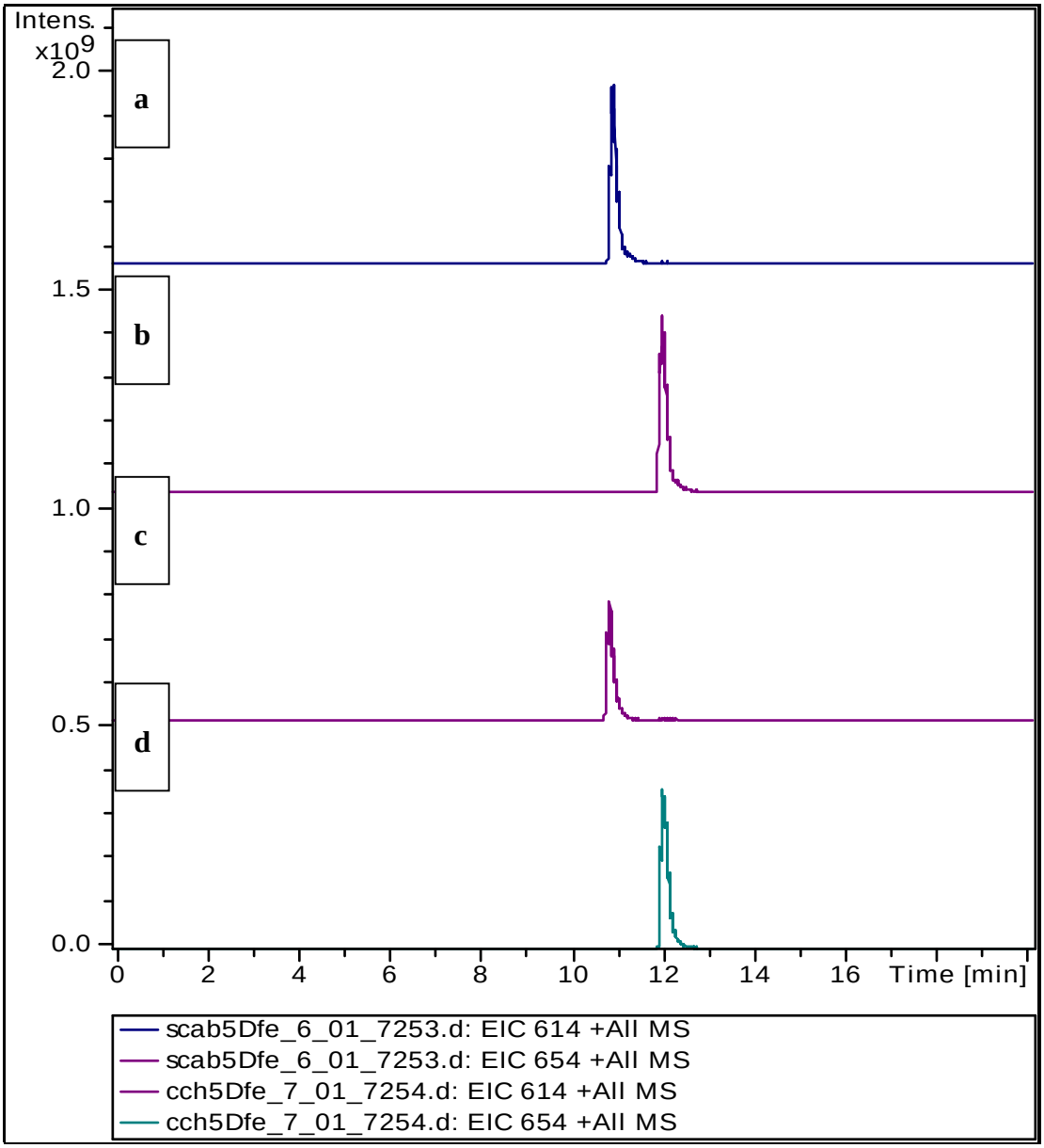


Figure 4-7 LC-MS confirming production of desferrioxamines in *S. scabies* 87.22. a) Elution peak of ferrioxmine B in *S. scabies* 87.22. b) Elution peak of ferrioxmine E in *S. scabies* 87.22. For comparison, c) ferrioxamine B peak from *"S. coelicolor"* A3(2), d) ferrioxamine E in "*S. coelicolor*" A3(2). This analysis is part of the genome project collaboration undertaken by G. L. Challis and L. Song and kindly provided for illustration in this work. Mass of compounds has also confirmed by HR-TOF MS-MS (not shown).

Ectoines

Genes closely related to those known to encode enzymes for biosynthesis of ectoines were located SCAB70711-SCAB70751. These gene clusters are very similar in "*S. coelicolor"* A3(2) and *S. avermitilis* MA-4680 by blast comparison; genes are also present in *S. griseus* and *S. chrysomallus* (Bursy *et al.* 2008) hence could be widespread in the *Streptomyces* genus. The importance of ectoine and 5-

hydroxyectoine biosynthesis in *"S. coelicolor"* A3(2) as a response to heat and salt stress is demonstrated in the same work (Bursy *et al.* 2008). The importance of these compatible solutes is likely to be similar in *S. scabies* 87.22 for avoiding dessication, heat or salt stress during pathogenicity or otherwise.

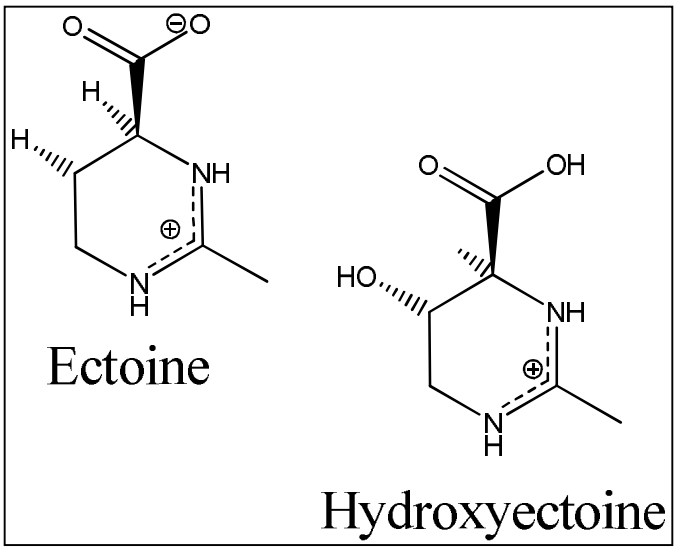


Figure 4-8 Molecular structures of ectoine and derivative hydroxyectoine. (After Bursy *et al.* 2008.)

4.3 Summary

4.3.1 Genome overview

The huge bacterial genome of *S. scabies* 87.22 has features in common with the other published complete streptomycete genomes, as would be expected, despite the extra niche available to *S. scabies* 87.22 as an opportunistic plant pathogen. Obligate pathogens appear to have undergone genome size reduction by loss of nonessential functions compared to related non-pathogenic organisms (Bentley and Parkhill 2004). *S. scabies* 87.22 has a still larger genome than the non-pathogenic streptomycetes which makes sense - as an opportunist *S. scabies* is able to carry all the genetic material for saprophytic growth as well as for pathogenicity.

The existence of saprophytic reproduction in *S. scabies* 87.22 has significance for agricultural crop losses; since the organism is reproducing in soil, removal of dead plant material assumed to be harbouring spores is not sufficient to ensure reduce

numbers of scab-affected tubers. Features of the genome likely to be directly involved in pathogenicity have been studied in some detail, and are summarized in **Chapter 5**.

4.3.2 Conserved gene clusters

Half of the likely complex natural product biosynthesis clusters judged in this work as not conserved in the other two strains, but these cluster genes are found amongst the third of *S. scabies* 87.22 coding sequences not conserved by the methods used above. This group of coding sequences has significance for the function of *S. scabies* 87.22 because biosynthesis of a complex natural product, the phytotoxin thaxtomin, is one known pathogenicity factor.

Another bioactive compound (concanamycin) is also known to be produced by this organism (Natsume *et al.* 2001) although the genes likely to encode production capacity in this organism have been identified for the first time in this work. Given the astonishingly wide capacity of organisms in this genus for biosynthesis of complex natural products, it is possible that other complex natural products with biological activity play a role in the pathogenicity of this organism. For this reason the DNA sequence of such clusters found in *S. scabies* 87.22 which are not also found in the other two available complete genome sequences have been studied in depth in this work and results of those studies are reported in **Chapter 6**.

Predictions from the presence of biosynthetic genes amongst those determined to be conserved across the three streptomycete genomes available for this study which have already been confirmed as natural products include germicidins and desferrioxamines (G. L. Challis and L. Song, pers. comm.); and hopene, additionally confirmed as non-essential for growth of *S. scabies* under a range of stress conditions in the laboratory (Seipke and Loria 2009).

Prediction from the genome sequence of the presence of gene clusters for biosynthesis of complex natural products is an important first step for deciphering the potential capabilities. Studies of much greater depth than is possible at a whole-genome level are necessary if the genetic basis of production is to be unravelled. Operators sequences have not been examined in this work with one exception (see **5.2.4**) due to the constraints of time. There is a great deal of potential for such

studies, and many further structural studies are possible; those lines of investigation may suggest testable hypotheses about why some clusters are cryptic in one organism and expressed in another, such as in the example of the clusters for melanins, apparently silent in *"S. coelicolor"* A3(2), but not in *S. scabies* or *S. avermitilis* MA-4680.

It is possible that the gene clusters conserved across the three genomes (hopanoids; two conserved NIS systems; geosmin; carotenoids; desferrioxamines; melanins; ectoines – see Table 4-6) are conserved across the *Streptomyces* genus. Although the environmental role of some of these biosynthetic products is not clear, they may have functions in the niches of these soil organisms which results in their maintenance by natural selection.

4.3.3 Method evaluation

OrthoMCL

OrthoMCL (Li, L. *et al.* 2003) results described in **4.2.2.6** were not entirely satisfactory because, for example, the similarities between non-ribosomal peptide synthetase systems led to several NRPS systems being assigned as orthologs when the products of these systems are demonstrably different. Another method was not attempted in this work as another researcher associated with the genome project has assigned orthologs using the reciprocal fasta (Pearson and Lipman 1988) method so there seemed little point in duplicating labour. OrthoMCL is designed for eukaryotic systems (Li, L. *et al.* 2003) and the differences between bacterial and eukaryotic systems may account for the less useful results in this organism.

Bacterial transcriptional activator domains PF03704

The results of a search for predicted proteins carrying the domain PF3704 described in **4.2.2.9** could be extended. AfsS in *"S. coelicolor"* A3(2) is small protein, just 63 residues, and it could have been missed by the Glimmer 3 automated prediction algorithm used on the complete genome sequence, which is better at predicting large coding sequences from the way the search models are built. It might also have fallen below threshold values for the `tblastx` overlay used to control for false negative predictions during curation. Given this bias it would make sense in further

investigations of BTAD-related signalling cascades in *S. scabies* 87.22 to check regions of DNA around BTAD regulators for AfsS homologues. This could be done by direct similarity searching such as blast, by searching for variations on the amino acid motif **TxxDHMPxxPA** reported to appear three times in the AfsS primary sequence (Martin 2004), or perhaps by a conserved domain search using the `Pfam B` domain which matches the AfsS sequence, PB163911.

Version control

The space requirements for longterm storage of the various versions produced by integrating each layer of sequence decoration into the main file are not great. Although it was not called on during annotation, it is important to keep a version from each stage of data construction so it is possible to remove the effect of any particular set of information if it is found to be unhelpful.

5 Results – sequences involved in pathogenicity

5.1 Introduction

5.1.1 Virulence factors in *S. scabies*

Pathogenicity in *S. scabies* is known to involve several factors. Production of phytotoxin thaxtomins is associated with the ability to cause scab disease (Loria *et al.* 2008) and several scab-causing streptomycetes have been confirmed to have a highly conserved region around the *nec1* locus including a putative transposase which may aid mobility of these sequences (Bukhalid and Loria 1997) as is more fully described in **1.3 *Streptomyces scabies* (or scabiei)** *the plant pathogen.*

A large pathogenicity island (PAI) was identified in the related scab-causing organism *S. turgidiscabies* Car8 and partially sequenced (Kers *et al.* 2005), and this has been used to identify conserved regions in the *S. scabies* 87.22 genome which may be involved in pathogenicity. In many pathogens iron is a crucial factor and virulence factors often have iron-dependent activity, which may be activated by low iron conditions inside the host organism. Since virulence traits in *S. scabies* 87.22 are known to be at least partially mobile, a survey of potentially mobile regions in the *S. scabies* 87.22 genome is included here to identify regions of difference with the non-pathogenic streptomycetes, *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680.

5.1.2 Iron-dependent regulation of pathogenicity traits

Several pathogens are known to show iron-dependent control of pathogenicity traits, for example the causative agent of diphtheria, *Corynebacterium diphtheriae* (Yellaboina *et al.* 2004a). DtxR, the diphtheria toxin repressor has been the subject of many studies and has been found to be a metal-ion binding protein. Lack of iron relieves repression of iron uptake and toxin production genes by causing this protein to release its binding site on the bacterial chromosome, revealing the -35 promoter region for these genes. The operator sequence bound by DtxR-like proteins and overlapping the -35 region has a conserved structure across several organisms. Related proteins have been identified in other organisms (Wennerhold and Bott 2006); (Wisedchaisri *et al.* 2004); (Boland and Meijer 2000). Similar sequences have

been identified in streptomycetes including the operator sequences for production of the desferrioxamine siderophore (Flores and Martin 2004; Tunca *et al.* 2007).

5.2 Results and discussion

5.2.1 Method development

Since pathogenicity traits are known to be transferable by mating between *Streptomyces* strains (Kers *et al.* 2005), this analysis began with listing regions in *S. scabies* 87.22 which are not found in either of "*S. coelicolor"* A3(2) or *S. avermitilis* MA-4680, and are over 10 k base pairs in length (Table 5-1 Insertion or deletion regions >10 k base pairs by comparison between streptomycete genomes, including pathogenicity loci in *S. scabies* 87.22. ***S. scabies* 87.22 compared to "*S. coelicolor*" A3(2) and *S. avermitilis* MA-4680. Table is ordered by presence of tRNA and int genes, then sequence order. Entries highlighted in yellow are those upon which known pathogenicity-associated features were found. RD= "region of difference".**Table 5-1). These may be mobile elements, and were subsequently inspected for conserved genetic material associated with horizontal transfer. The `alien_hunter` script was used to generate one set of lists of possibly mobile elements, by variable order composition deviation signatures and the presence of known insertion sites (Vernikos and Parkhill 2006). The list generated using `alien_hunter` was very long and included some false positives when checked, possibly due to compositional bias of highly conserved sequences for essential functions (G. S. Vernikos pers. comm.).

Hence, a shorter list of candidate mobile elements were identified by eye from gaps in conserved sequence order by `tblastx` comparison of three streptomycete genomes visualized in `ACT` (See 3.1.1 for details). The list presented here (Table 5-1) was compiled using both sources. A subsequent investigation with special interest in potentially mobile regions could well use a more systematic approach and work through the list of `alien_hunter` candidate regions, eliminating false positives.

Due to lack of synteny in the arm regions, it was not possible to identify mobile elements in *S. scabies* 87.22 by the comparison method before 2.4Mth base pair or beyond 8.85 Mth base pair. The most significant insertion or deletion regions greater

than 10 kbp in size are listed in Table 5-1 below. The size of the region in base pairs, and presence or absence of tRNA sequences indicating possible site-specific insertion points, are listed in the table, as well as the presence or absence of possibly functional integrase domains which may indicate capability for transfer (Integrase domain and tRNA tab files from output of alien_hunter (Vernikos and Parkhill 2006).

RD #	start	end	size	tRNA?	int?	note
5	2400974	2610926	209952	yes	yes	mosaic feature several transposase and int matches.
8	3304766	3349784	45018	yes	yes	integrated plasmid? spdABCD; excisionase.
21	5253507	5275621	22114	yes	yes	integrated plasmid? spdAB, rep.
22	5383635	5462369	78734	yes	yes	prophage? Two pairs of repeated tRNAs nearby
23	5825778	5866847	41069	yes	yes	contains mobilisation relaxase
26	6256425	6273761	17336	yes	yes	spdB; traB; xis; int;
29	6736508	6797747	61239	yes	yes	prophage?
2	1780895	1856225	75330	yes	no	indel vs SCO and SAV; contains 2 alien predictions
14	4506051	4591103	85052	yes	no	transposase
16	4671887	4685821	13934	yes	no	inverted sections in both sco and sav here
19	4813578	4881377	67799	yes	no	transposase features - conserved spore pigment cluster
30	6929793	7024806	95013	yes	no	no obvious mobility features; hybrid cluster, tRNA-6929739
9	3595928	3774724	178796	no	yes	many mobility features. txt; conservon; lantibiotic
11	3981729	4003871	22142	no	yes	transposase, extracellular prot inc SCO6220-homolog
15	4614026	4638408	24382	no	yes	integrases, cargo of hypothetical proteins
20	4884562	4987353	102791	no	yes	transposase; cargo includes hybrid PKS protein
24	5924891	5949920	25029	no	yes	inverted vs SCO and SAV; contains transposase
25	5953988	5970190	16202	no	yes	indel vs SCO and SAV; alien prediction goes further 5'
31	7155474	7203346	47872	no	yes	prophage?
34	8472686	8578879	106193	no	yes	three transposase elements plus nec1
1	1543760	1715443	171683	no	no	several transposase-like features
3	2104893	2167022	62129	no	no	
4	2292641	2329739	37098	no	no	
6	2640432	2707743	67311	no	no	
7	3132697	3158295	25598	no	no	
10	3837615	3848957	11342	no	no	
12	4160484	4175309	14825	no	no	
13	4236348	4248986	12638	no	no	
17	4694987	4711027	16040	no	no	
18	4770615	4789517	18902	no	no	
27	6304373	6321871	17498	no	no	tRNA just outside 3' end
28	6503000	6531432	28432	no	no	
32	7263735	7314362	50627	no	no	
33	7730084	7743937	13853	no	no	
35	8700414	8845245	144831	no	no	carries cfa-derivative and hybrid cluster

Table 5-1 Insertion or deletion regions >10 k base pairs by comparison between streptomycete genomes, including pathogenicity loci in *S. scabies* 87.22. *S. scabies* 87.22 compared to "*S. coelicolor*" A3(2) and *S. avermitilis* MA-4680. Table is ordered by presence of tRNA and int genes, then sequence order. Entries highlighted in yellow are those upon which known pathogenicity-associated features were found. RD= "region of difference".

It is more likely by parsimony that these regions are insertion in the *S. scabies* 87.22 lineage than deletions in identical positions in the other two lineages. The possibility that some factor controls the likelihood of deletion and loss of certain sections of

sequence cannot be ruled out however, hence the use of ‘insertions or deletions’ to refer to these regions of difference.

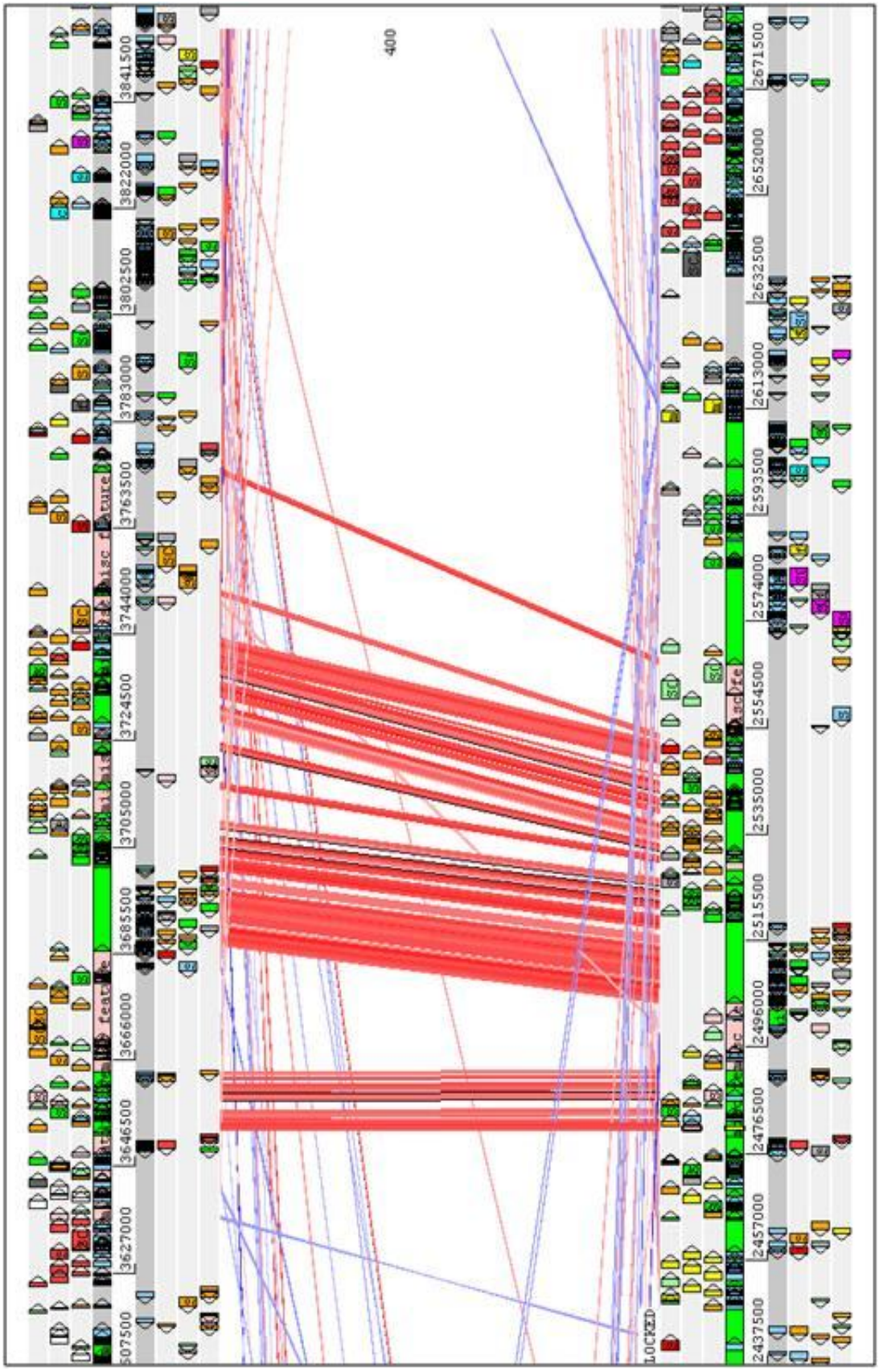


Figure 5-1 Self-match blastn comparison visualised in ACT between the duplicated regions of *S. scabies* 87-22. Above, RD9 (Table 5-1), which is associated with pathogenicity island features including *txt* gene cluster. Below, RD5(Table 5-1).

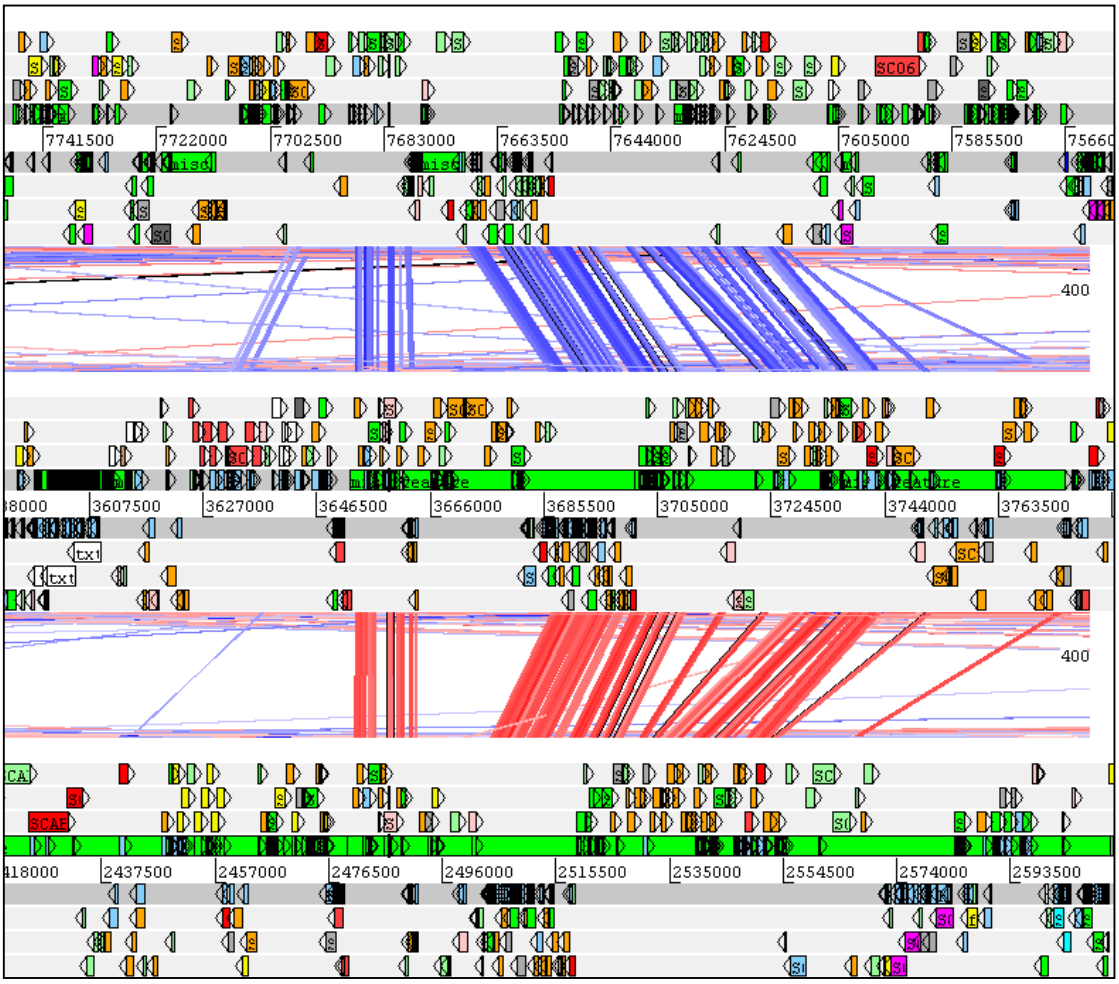


Figure 5-2 Comparison of duplicated region in *S. scabies* 87-22 and matching region in "*S. coelicolor*" using `tblastx` visualized in `ACT`. Matching regions aligned (top to bottom) matching region in "*S. coelicolor*" A3(2); RD2 copy in *S. scabies* 87-22 associated with pathogenicity features; RD1 copy in *S. scabies* 87-22. From this comparison, the copy of the duplicated region inside the first candidate mobile feature (RD1) is the one most closely related to the copy in "*S. coelicolor*" A3(2).

Two regions of difference (Table 5-1, RD5 and RD9) contain apparent duplications of a sequence region between 70 (RD5) and 90kbp (RD9) in size (**Error! Reference source not found.**). These regions became obvious from use of the `TribeMCL /cluster` qualifier, (see Chapter 2 for details of method) and from self-match data from `blastn` comparison of the *S. scabies* 87.22 genome against itself. These regions of the *S. scabies* 87.22 genome show broadly conserved gene order with several insertions or deletions. These two regions could encode mobile elements related to each other and diverged since duplication.

`ACT` comparison between the three genomes showed further that a region of the "*S. coelicolor*" A3(2) genome near the 7683000th base pair had similarity to the

duplicated regions (Figure 5-2) . This region is not in the general trend of conserved gene order between the three genomes, and may represent a rearrangement or a mobile region integrated into *"S. coelicolor"* A3(2) as well. It appears that the region in *"S. coelicolor"* A3(2) is most closely related to the copy of the duplicated region contained in RD5.

5.2.2 PAI fragments in the *S. scabies* 87.22 genome

Known pathogenicity genes in *S. scabies* 87.22 by comparison with the pathogenicity island sequence from *S. turgidiscabies* Car8 are found in at least two locations on the genome, near 3.6 Mth base pair (Table 5-1 RD9) and near 8.5 Mth base pair (Table 5-1 RD34). Genes for biosynthesis of the phytotoxin thaxtomin are found in one, and the *nec1* locus and saponinase genes in the other.

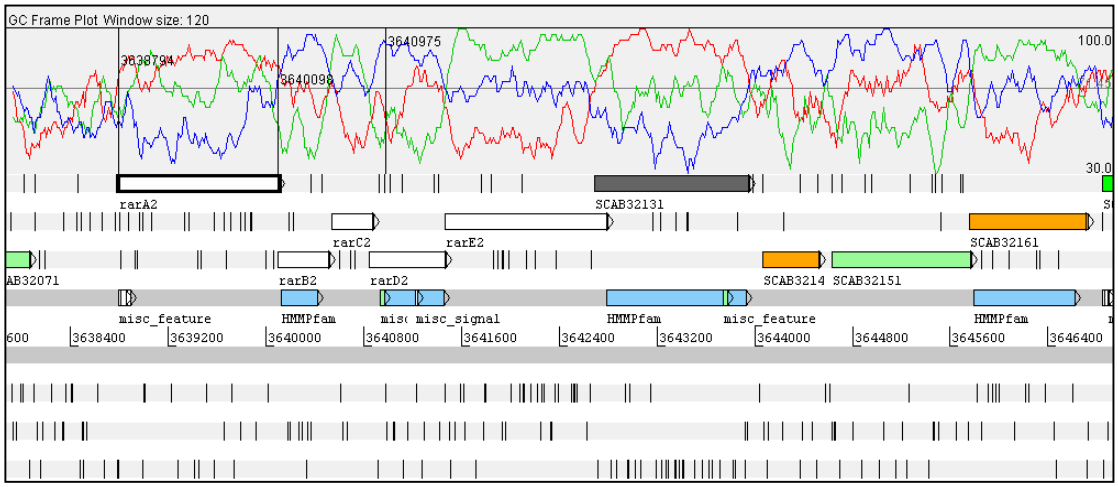


Figure 5-3 Conservon cluster in putative insertion region RD9 of *S. scabies* 87-22 associated with thaxtomin biosynthesis genes. Conservon genes rarA2B2C2D2E2 shown in white.

5.2.2.1 First PAI fragment, bases 3595928 to 3774724, RD9

Coding sequences for biosynthesis of thaxtomin are found in a region of the *S. scabies* 87.22 genome approximately between base pairs 3595928 and 3774724, marked as RD9 in Table 5-1. An insertion or deletion is apparent in the genome of *S. scabies* 87.22 here in comparison to *"S. coelicolor"* A3(2), *S. avermitilis* MA-4680 (Figure 5-4). A cluster proposed to encode the primary sequence and tailoring of a lantibiotic is also found in this region. Following the putative lantibiotic cluster is a conservon (Figure 5-3).

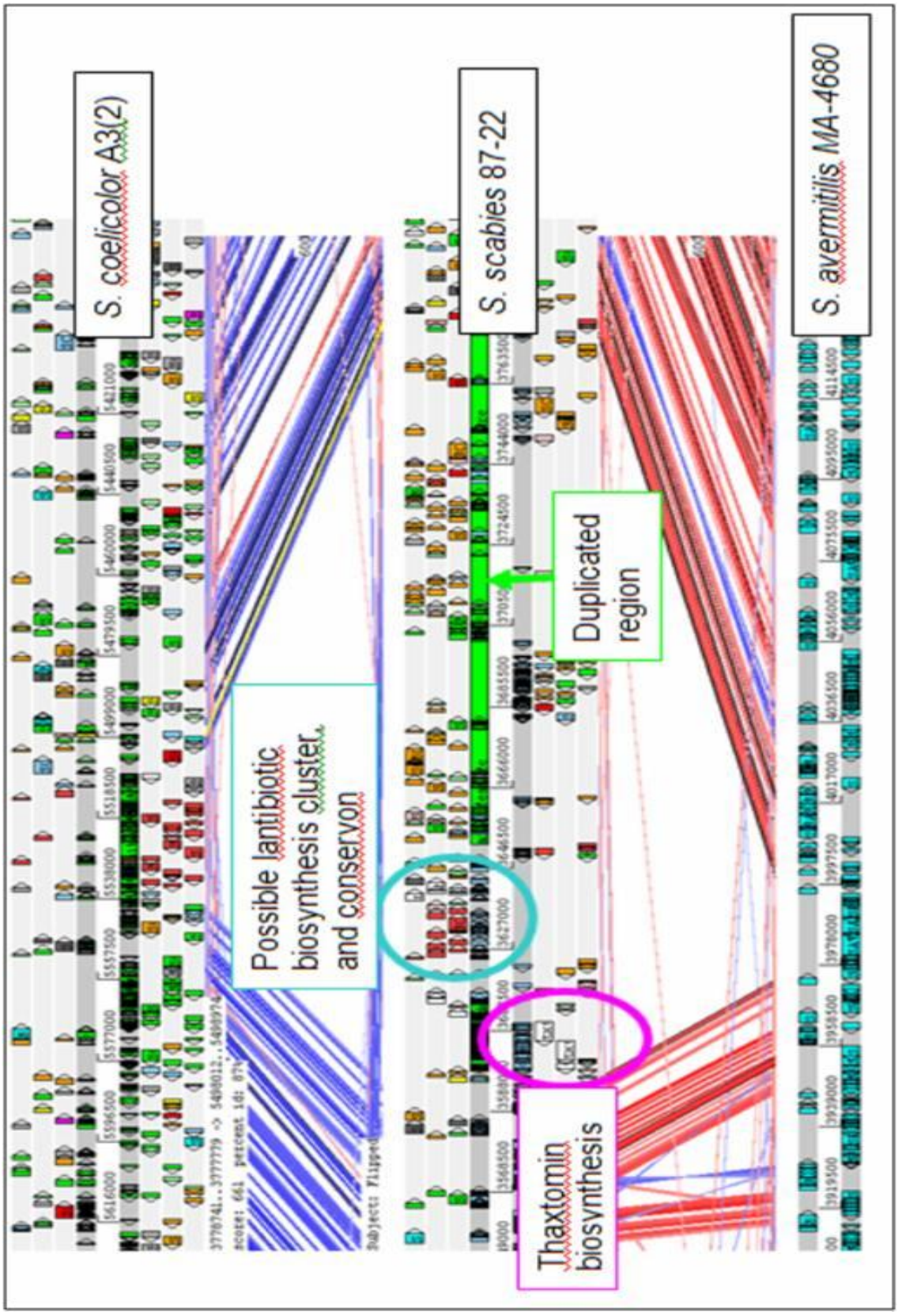


Figure 5-4 Pathogenicity genes in *S. scabies* 87-22 genome (1) txt and nearby sequences visualized in ACT. Comparison with `tblastx` between (top) *"S. coelicolor"* A3(2), (middle) *S. scabies* 87-22, and (bottom) *S. avermitilis* MA-4680. Gap shown here where there are no matches to *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680 is proposed to be an insertion in *S. scabies* 87-22.

Green feature shown on the nucleotide sequence line of *S. scabies* 87-22 indicates the duplicated region.

Conservons (Figure 5-3), first identified in *"S. coelicolor"* A3(2), may encode a membrane-associated signalling heterocomplex with some similarity to the eukaryotic G-coupled receptor system (Komatsu *et al.* 2006). It is possible that this gene cluster encodes a signalling complex associated with pathogenicity and thaxtomin production, because it is next to the *txt* cluster in region of difference RD9 (Figure 5-4). Coding sequence *rarD2* contains a TTA codon, which could indicate BldA regulation if that system operates in *S. scabies* 87.22. The aligned region with RD9 in *"S. coelicolor"* A3(2) contains the gene cluster for biosynthesis of actinorhodin, and another complex biosynthesis cluster is found in *S. avermitilis* MA-4680 at the same point. Several conservons in S. *scabies* 87.22 have TTA codons indicating the possibility of regulation by availability of TTA-Leu tRNA encoded by *bldA*.

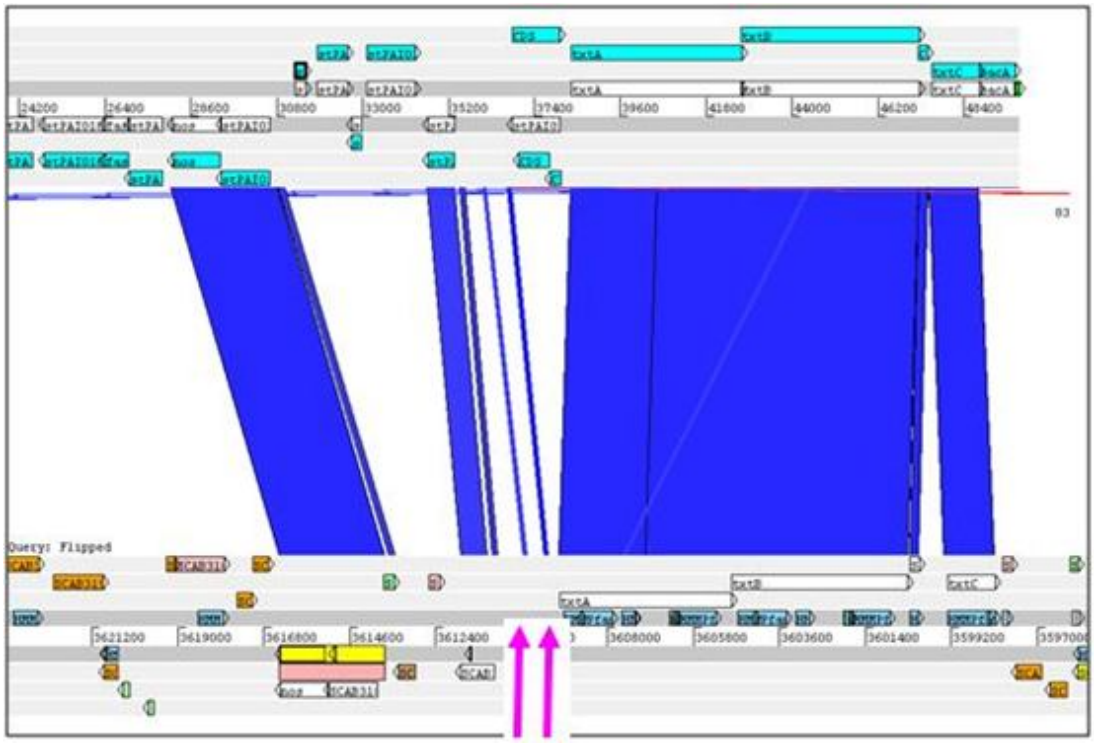


Figure 5-5 Comparison (blastn, visualized in ACT) between section of *S. turgidiscabies* Car8 PAI sequence (top) and *S. scabies* 87-22 RD9. Several insertions and deletions can be seen in the comparison, as well as a conserved region of intergenic sequence which may contain operator sequences for control of gene expression (indicated with magenta arrows).

Genes for biosynthesis of thaxtomin are encoded on this fragment of the pathogenicity island (Figure 5-4 above), and these are discussed in more detail in Chapter 6, **6.2.3.1**. Several insertions or deletions are apparent in this section in comparison of the *S. turgidiscabies* Car8 PAI section and RD9 in *S. scabies* 87.22. An intergenic region shows clear similarity (magenta arrows in Figure 5-5) despite apparent insertions and deletions in this region of sequence. This region contains a sequence similar to the binding site for an IdeR-like 'iron-box' repressor which are described further in 5.2.4 below, and could be a target for investigation of the molecular control of pathogenicity initiation.

Several coding sequences found in the contig containing this section of the *S. turgidiscabies* Car8 pathogenicity island are not found in *S. scabies* 87.22 and vice versa. These coding sequences hence might not be essential for pathogenicity. It cannot be concluded that they are definitely not involved in pathogenicity – it might be the case that an essential function of one of those coding sequences in one of the organisms is complemented by a different coding sequence in the other.

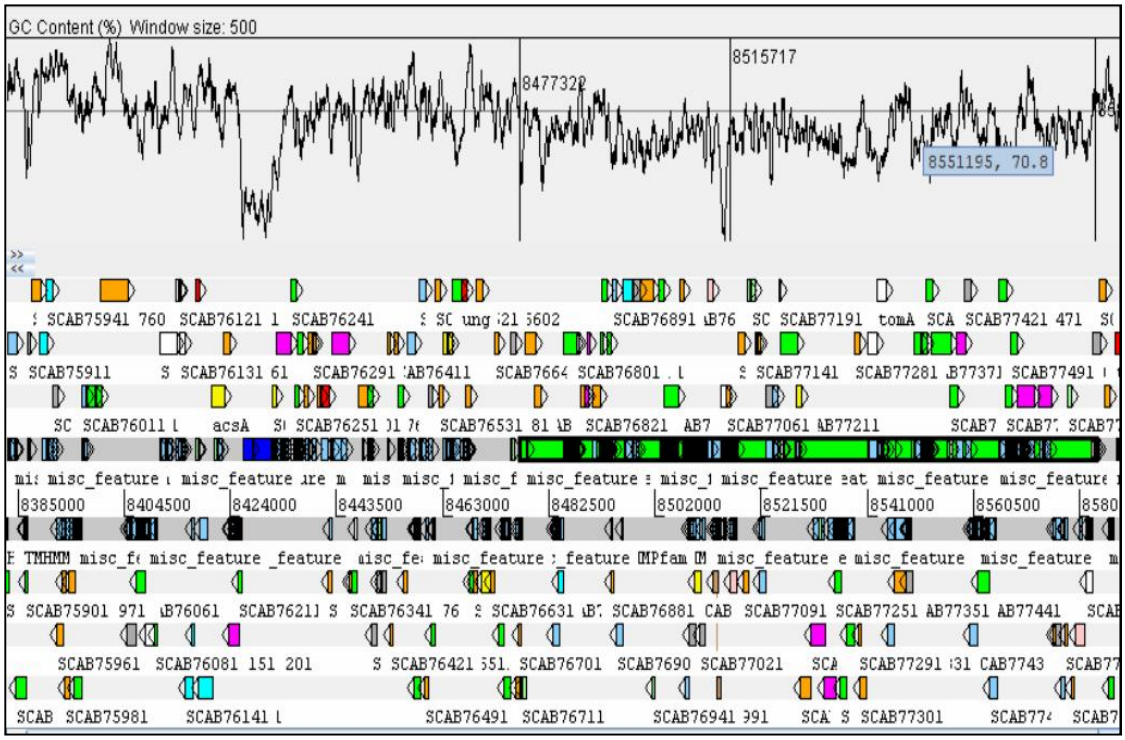


Figure 5-6 Artemis overview of second PAI fragment in *S. scabies* 87-22 showing G+C content deviation associated with nec1 region. Green feature represents RD34, the proposed region of the integrated PAI.

5.2.2.2 Second PAI fragment, 8472686..8578879, RD34

This second region of sequence with similarity to the *S. turgidiscabies* str. Car8 PAI surrounds the necrogenic *nec1* locus. A section of sequence here is identical over 7672 base pairs, which probably encodes seven coding sequences (Figure 5-8), including *nec1*. The absence of stPAI006 in *S. scabies* 87.22 suggests it is an insertion sequence in *S. turgidiscabies* Car8. The ORFtnp transposase-like coding sequence is found with a frameshift in both strains. This may indicate that it has been inactivated, and perhaps the mobility of this section of DNA is catalysed by the coding sequences SCAB77071 or SCAB77031 which appear to have horizontal-transfer-associated domains.

In comparison with the non-pathogenic streptomycetes, there are alternative insertions at this point in the genomes of *S. avermitilis* MA-4680 and "*S. coelicolor*" A3(2), suggesting that some feature of the conserved boundaries of the insertions favours integration of novel genetic material at this point (Figure 5-7).

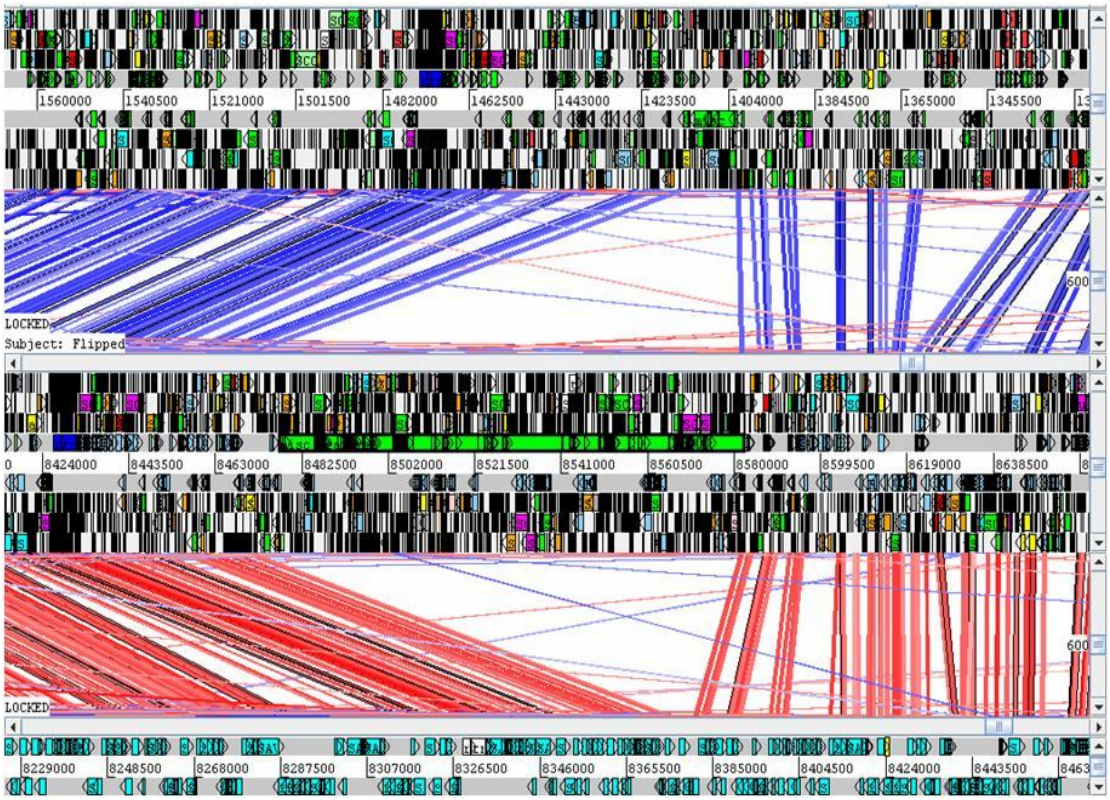


Figure 5-7 Comparison (`tblastx` visualised in ACT) showing second pathogenicity-associated insertion. *"S. coelicolor"* A3(2), *S. scabies* 87.22 and *S. avermitilis* MA-4680. This region contains the necrosis factor *nec1*, a putative tomatinase gene, and transposase-like sequences previously found in associated with these pathogenicity factors.

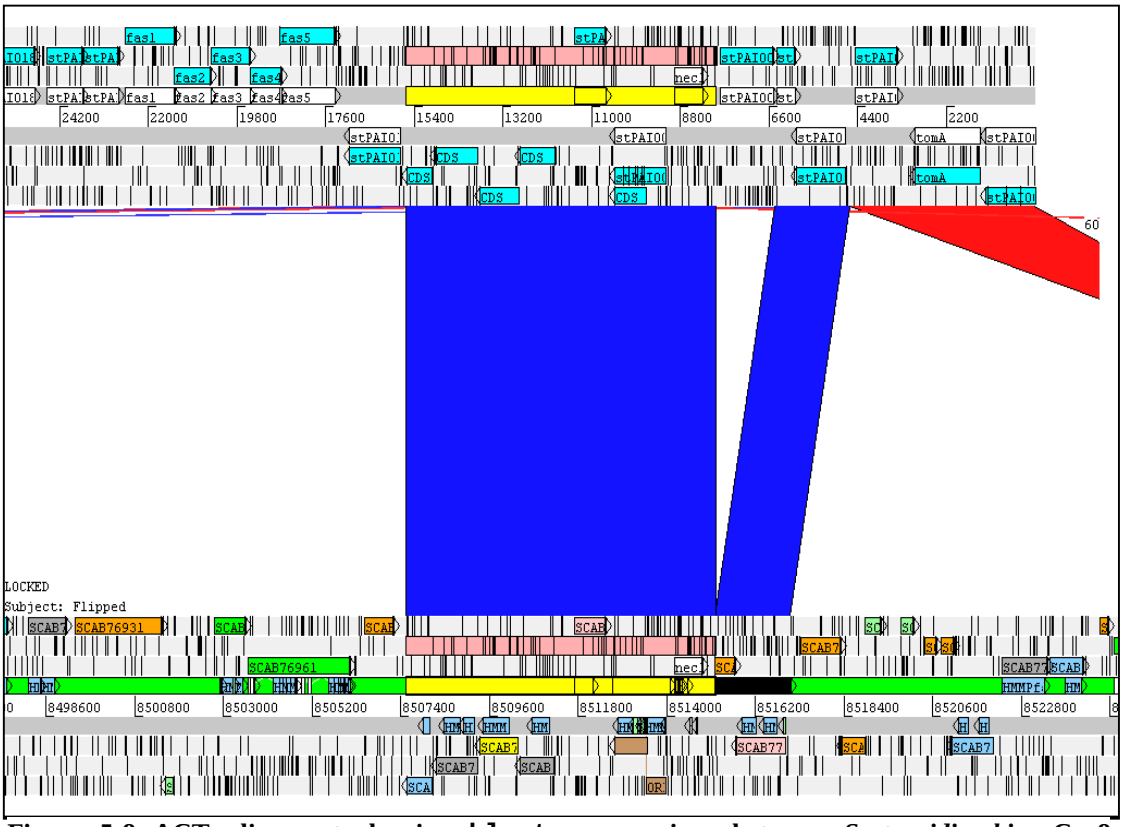


Figure 5-8 ACT alignment showing `blastn` comparison between *S. turgidiscabies* Car8 pathogenicity island (top) and related PAI fragment RD34 region of *S. scabies* 87-22 genome (bottom).

This region of sequence in *S. scabies* 87.22 is found to contain the putative saponinase also found in the *S. turgidiscabies* Car8 PAI sequence (Figure 5-9). This fragment is reversed in orientation in *S. scabies* 87.22 compared to the *S. turgidiscabies* Car8 PAI assembly (Figure 5-8). This region of the *S. turgidiscabies* Car8 PAI is a separate contig to the neighbouring fragment containing the *nec1* region, and hence this may represent a misassembly in the PAI sequence. If not, the reorientation of this saponinase coding sequence and its neighbours in relation to the *nec1* region may have been catalysed during one of the recombination events.

The two regions so far identified (RD9 and RD34) contain 284 989 base pairs, which although large is still somewhat short of the approximately 660 kbp pathogenicity island reported in *S. turgidiscabies* (Kers *et al.* 2005). A fasciation operon is known to be encoded within the pathogenicity island in *S. turgidiscabies* Car8, thought to be responsible for gall formation, as is the co-linear operon found in the plant pathogen *Rhodococcus fascians* (Crespi et al. 1994). This *fas* operon is not present in the genome of *S. scabies* 87.22, accounting for approximately 12 kbp of missing sequence.

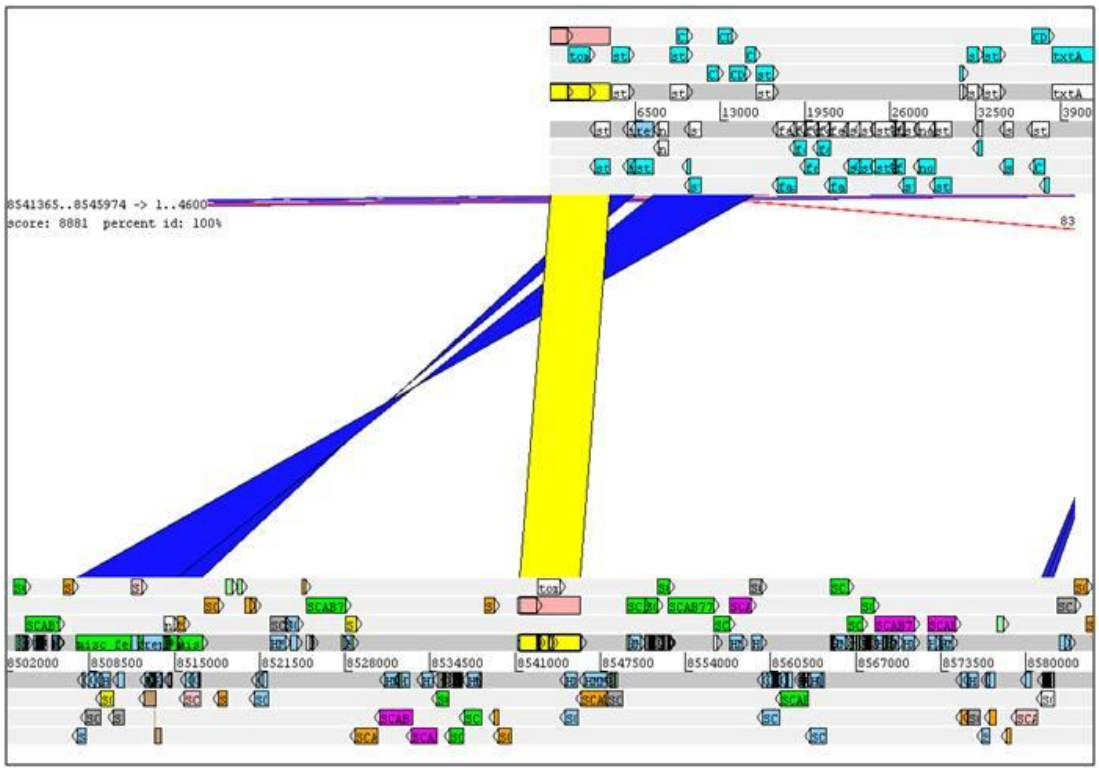


Figure 5-9 `Blastn` comparison of right hand end of *S. turgidiscabies* Car8 PAI (top) with RD9 of *S. scabies* 87-22 genome, illustrating relative positions of putative saponinase coding sequence (highlighted in yellow) and the nec1 region (blue inverted match).

The part of the pathogenicity island found with the thaxtomin biosynthesis cluster in RD9, towards the left arm of the *S. scabies* 87.22 genome, probably originally inserted near the right core/arm boundary as part of RD34. This section may have has subsequently moved to its present position by duplication using homologous recombination which has been studied in bacterial systems (Shen and Huang 1986; Mahan and Roth 1988; Mahan and Roth 1989).

There are several sites at which several hundred base pairs are exactly duplicated in the *S. scabies* 87.22 genome. The hypothesis of the original insertion of a pathogenicity island at the right core/arm boundary where this second region bearing pathogenicity traits is still found, is supported by the fragments of the *bacA*-like gene, which is a known insertion site for the pathogenicity island (Kers *et al.* 2005). Fragments of this gene are found at both ends of the insertion region containing the *nec1* locus, supporting the suggesting that the insertion target is the same in *S. scabies* 87.22.

Four identical repeats of 771 bp were identified (see section in Methods) at base positions 18039..18810, 3744492..3745254, 8894945..8895716, and

10130657..10131429. Two of these, the first and last, are in the terminal inverted repeat region. The other two provide a possible explanation for the divided pathogenicity genes in *S. scabies* 87.22 (Figure 5-10); the identical sequence may have provided a position for recombination by arm exchange during replication.

If the insertion or deletions between position 8472686 and 8895716 were considered to be one section (including RD34, a large region common to the non-pathogenic streptomycetes as well, and RD35) the total size of that region would be 423 030 bp. The other section known to be part of the *S. turgidiscabies* Car8 PAI is 178 796 bp long (RD9). Adding the two gives a total of 601 826 bp which is nearer to the

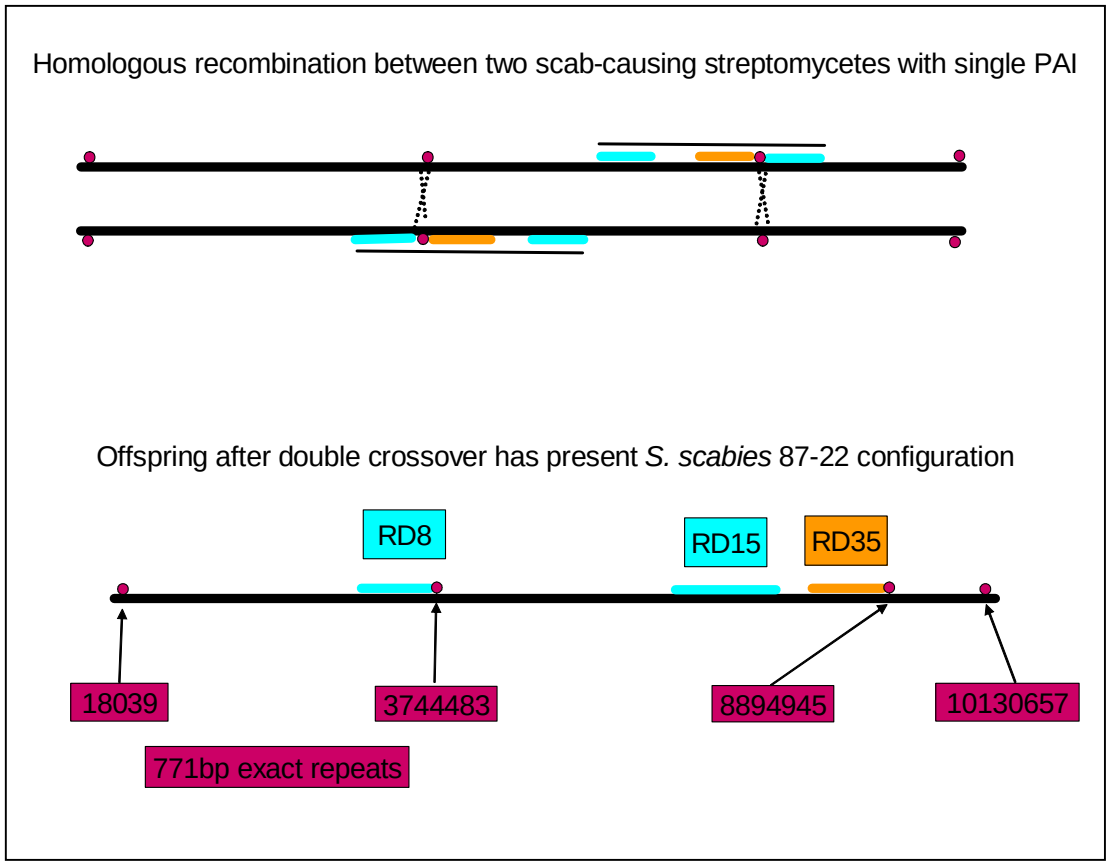


Figure 5-10 Possible recombination mechanism for splitting pathogenicity island sequences in *S. scabies* 87-22 using 771 base pair exact repeats.

estimated size of the largest transferred PAI sections (Kers *et al.* 2005). As streptomycete DNA has very high G+C content, it runs lighter than DNA with a less biased DNA content, and this can have a very large effect on size estimates: the whole chromosome of *S. scabies* 87.22 was estimated to be 8.5 Mbp by mobility on gel, but the complete sequence is over 10.1 Mbp. Hence the pathogenicity island from *S. turgidiscabies* Car8 might be larger than 660 bp when completely sequenced.

RD35, which may if this mechanism is correct be found to be part of the pathogenicity island, represents a third large region of difference in ACT comparison of the synteny of the three available genomes (*S. scabies* 87.22, *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680). This region is found at 8700414-8845245 bp in *S. scabies* 87.22, and represents a possible insertion or deletion of over 144 kbp. It carries a hybrid PKS/NRPS cluster (for more details see Chapter **6.2.5.4**), another conservon, and the polyketide synthase gene cluster which probably produces a compound related to coronafacic acid **6.2.5.1**. This cluster may be involved in virulence since in *Pseudomonas syringae* coronafacic acid is a precursor of the phytotoxin coronatine (Liyanage et al. 1995) and is discussed in more detail in Chapter 6.

5.2.3 Genes not on PAI and possibly involved in pathogenicity

5.2.3.1 Secreted esterase

Esterase A, characterised in a strain of *S. scabies* (Raymer *et al.* 1990) (see further Introduction **1.3.3**) appears most similar to SCAB3021, but this is not a striking similarity, with only 28% of bases identical by gapped blast. Predicted protein SCAB3021 has no above-threshold conserved domains using Pfam; subthreshold Pfam domains PF00657 (GDSL-like Lipase/Acylhydrolase, expect 0.28) and domain of unknown function #77 PF1910 (expect 0.98) were found.

Several possible positions for the start site exist, and the Kyte-Doolittle hydrophobicity plot does not have a very clear indication for the N-terminal hydrophobic region of the secretion signal. The start site chosen is the one that appears most correct by Frame plot (Bibb *et al.* 1984). A putative active site motif of the form **GDSYT** is found with proposed active site Ser at an appropriate position by comparison with the enzyme characterised in *S. diastatochromogenes* (Tesch *et al.* 1996) and in *S. scabies*, from which a model has been proposed from crystal structure (Wei *et al.* 1995).

There are a very large number of coding sequences that might encode secreted degradative enzymes in the genome; 479 coding sequences were coded as involved in degradation of large molecules. An additional number of coding sequences may encode secreted catabolic enzymes, but conserved domains or high levels of

similarity to characterised enzymes were not found during curation of the genome and hence these were annotated as secreted proteins (1466 identified as secreted by SignalP 2.0 (Nielsen *et al.* 1999); for more details see **2.5.5**.)

Since previous investigators have suggested that secreted esterases may play a part in pathogenicity, several others identified in this genome could be targets for further investigation besides the CDS SCAB3021. SCAB78931 encodes a secreted protein with similarity to cutinases, which are α/β hydrolases and virulence factors responsible for digesting the cuticle of plant cells by some fungal plant pathogens (Sweigard *et al.* 1992). SCAB78951 also appears to encode an extracellular esterase, and has some similarity to a mycobacterial antigen. Both these coding sequences are associated with the hybrid PKS/NRPS system in RD35 (described further in Chapter 6).

5.2.3.2 RTX toxin homologoues

This family includes hemolysins in *Escherichia coli* and some other known virulence factors (Kuhnert *et al.* 1997). In *Erwinia chrysanthemi* members of this family function as extracellular proteases (Delepelaire and Wandersman 1989). These proteins are found in an operon of form *rtxCABD*, where the A proteins encode the structural component, C proteins are acyl transferases responsible for acylation of a conserved lysine residue with a fatty acid to activate the structural component, B and D are components of the export system (Lally *et al.* 1999).

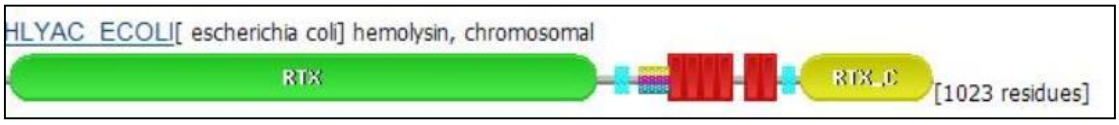


Figure 5-11 'Bead on a string' graphic of conserved domains expected in RTX toxin genes. From Pfam(Finn *et al.* 2006)

Three proteins in the *S. scabies* 87.22 genome have N-terminal secretion signals and contain the calcium-binding repeats found in RTX/hemolysin family proteins: SCAB6751, SCAB19481, SCAB63831. However, there are no conserved N or C terminal regions matching PF02382 (RTX_N) and PF08339 (RTX_C), nor are there any significant hits to the HlyC model (RtxC activating protein) PF02794, and these proteins are not in operons of the expected form. Hence it seems likely that these

proteins are a different type and there is no evidence to ascribe RTX toxin function to them.

5.2.3.3 Pectin breakdown capacity

Pectin is part of the primary wall of plant cells. There are several CDSs predicted to encode pectinesterases and pectate lyases in the *S. scabies* 87.22 genome (Table 5-2).

systematic identifier for coding sequence	product	similarity to characterised protein (by fasta)	conservation in other Streptomycete genomes	TAT signal?
SCAB44901	putative secreted pectate lyase	43% JC7653	SAV6375; SCO1880	yes
SCAB70521	putative secreted pectinesterase	44% aa_id BAB90989 (C term)	SAV6377; SCO1879 (partial)	no: mutation?
SCAB70551	putative secreted pectate lyase	42% aa_id JC7653	SAV6375; SCO1880	no
SCAB70561	putative secreted pectinesterase	37% aa_id BAB09089	SAV6376; SCO1879	yes
SCAB70571	putative secreted pectinesterase	35% aa_id BAB90989	SAV1377; SCO1879	yes
SCAB78781	putative secreted pectinesterase	41% aa_id BAB90989	SAV6376; SCO1879	yes
SCAB82041	putative secreted pectate lyase	33% aa_id BAA81753	SAV6382 (partial)	yes
SCAB82421	putative secreted pectate lyase	45% aa_id CAC33162	not conserved	no

Table 5-2 Coding sequences in *S. scabies* 87.22 with annotation as putative pectate lyase or pectinesterase.

The most closely related characterised protein to SCAB44901 is a thermostable enzyme with pectate lyase activity (pir||JC7653 pectate lyase (EC 4.2.2.2) PL47 - *Bacillus* sp. pdb|1VBL|A and pdb|2BSP|A). SCAB44901 has two twin arginine motifs near the start, which are necessary for export from the cell through the TAT pathway which is known to operate in at least some streptomycetes (Schaerlaekens *et al.* 2001). TAT secretion is likely to be necessary for this class of enzymes to function: similar enzymes, such as the one mentioned above which was studied in a *Bacillus* species above, binds calcium. Proteins such as these that bind metal ion cofactors usually need to be folded with the ion bound to be functional, and hence can only be exported through the TAT pathway. SCAB44901 has an imperfect repeat CGTTACGTAAGG upstream of the presumed coding start which might be a regulator binding site.

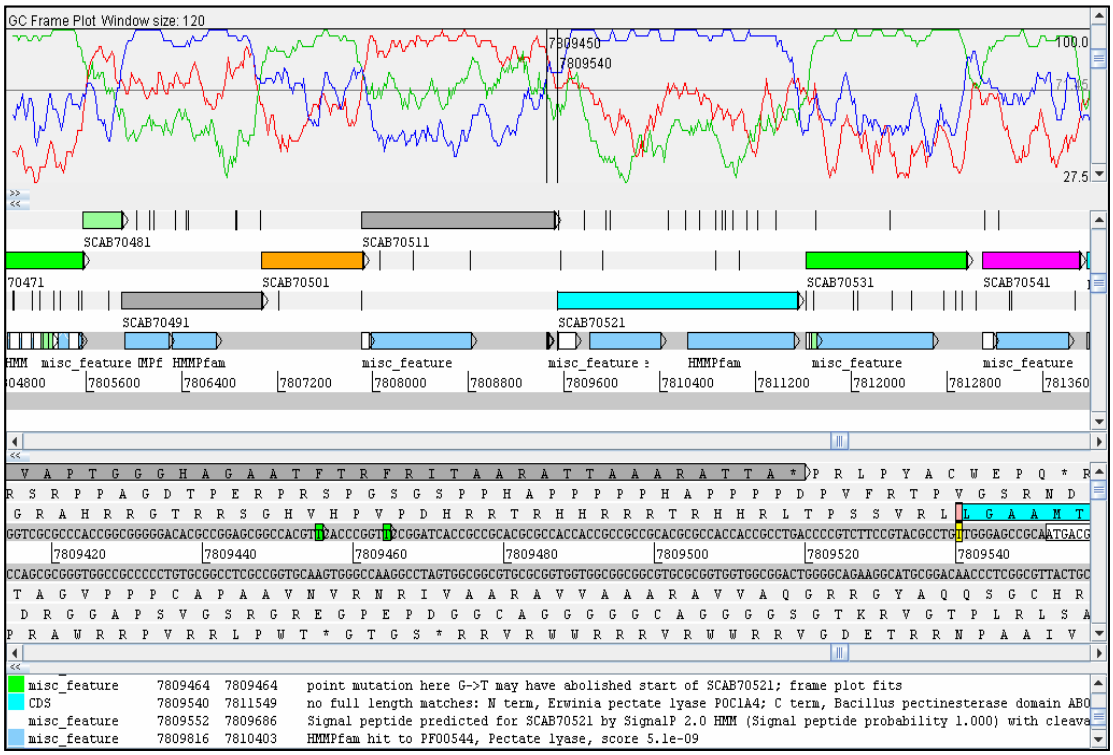


Figure 5-12 Artemis view of SCAB70521, showing frame plot and possible point mutations abolishing RR signal (green boxed bases).

The nearest characterised protein to SCAB70521 is a bifunctional pectate lyase/ pectinesterase from *Bacillus* sp. P-358. This is not a full length match; only the C terminal domain (matching PF01095) matches. The N terminal domain of SCAB70521 is most closely related to a characterised pectate lyase PelE from *Pectobacterium atrosepticum* and this domain is thought to have pectinesterase function.

The frame plot for SCAB70521 seems to indicate a start site further upstream than the one which has been selected, and it seems like that a point mutation at either of the positions marked (green boxed bases in close-up view in Figure 5-12) could have caused the translational start site to move downstream to the position at which it has been set. These mutations would have abolished the twin arginine signal motif which causes proteins to be directed to the TAT pathway. Proteins without the RR motif are unlikely to be functional for this reason, and this inactivation may have had a selective advantage.

If the coding sequences in this group with RR motifs for secretion through TAT pathway are considered, there is only one more than the number found in *S. avermitilis* MA-4680 and two more than in "*S. coelicolor*" A3(2). It is possible that these coding sequences are involved in pathogenicity, but scab disease doesn't have

the suite of pectin breakdown enzymes that soft rot pathogens such as *Erwinia* are known for. Pectin breakdown is unlikely to play a large part in pathogenicity in *S. scabies* 87.22 because soft rot is not known to be part of the phenotype.

5.2.4 Target sequences for iron-dependent repressor?

start	end	sense	sequence found	coding sequence	note
2180451	2180469		taaggtaagccttacctgg	SCAB19301	unchar. CHP
2180451	2180469	c	ccaggtaaggcttacctta	SCAB19291	unchar. GntR family regulator
2751075	2751093		tcaggttaggctcacctct	SCAB24341	sucrolytic enzyme, sensor-regulator pair?
2751075	2751093	c	agaggtgagcctaacctga	no CDS	-
5314653	5314671		ttaggtgaggctaacctaa	SCAB47411	peptidase family S45 (clan PB(S))?
5314653	5314671	c	ttaggttagcctcacctaa	SCAB47401	ViuB-like siderophore utilization?
6450935	6450953		ttaggttagcctaacctaa	SCAB57961	IS630 family transposase contains 2*TTA
6450935	6450953	c	ttaggttaggctaacctaa	SCAB57951	*desABCD* desferrioxamine biosynthesis
6454385	6454403		gcaggttagcctaacctca	no CDS	-
6454385	6454403	c	tgaggttaggctaacctgc	SCAB57981	*desFE* putative desferrioxamine export
7927932	7927950		agagggtaggctaacctaa	SCAB71661	iron-siderophore uptake? Like SCO1787
7927932	7927950	c	ttaggttagcctaccctct	no CDS	-
9433990	9434008		gtaggtgaggcttacctta	SCAB84491	lucA/C family protein, *pobA*
9433990	9434008	c	taaggtaagcctcacctac	SCAB84481	unchar. CHP, beta-fructosidase?
9569869	9569887		ctaggtaaggcttacctta	SCAB85521	peptide N-oxygenase? With *nrps5*
9569869	9569887	c	taaggtaagccttacctag	SCAB85511	formyltransferase? With *nrps5*

Table 5-3 Sequences identified by similarity to iron-dependent repressor-binding site. CHP=conserved hypothetical protein.

Several possible binding sites have been identified from their similarity to the target sequence of the iron-dependent repressors of the DtxR/IdeR family. Several of these are conserved positions by comparison on the alignment of *S. scabies* 87.22 with *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680. These conserved positions could be essential for general response to low iron conditions, and they include operator positions in the *des* gene cluster which encodes proteins for biosynthesis of desferrioxamine siderophores (SCAB57951, SCAB57981, Table 5-2). It is not clear what significance the presence of TTA codons in the insertion sequence SCAB57961 (and the other three copies of this sequence in the genome: cluster 0138 consists of SCAB35111, SCAB57961, SCAB59871, SCAB87221) has, since this rare codon is thought to be part of the *bldA* regulatory mechanism in *"S. coelicolor"* A3(2) (Hesketh *et al.* 2007).

This prediction of a possible regulon is speculative, generated from the sequence data alone: small changes in the search pattern make great differences in the number of target sequences identified. It may be that computational searches by base composition do not effectively replicate the sequence search performed by the

binding site of the enzyme, and the approach used is primitive in comparison to position weighted matrix approaches such as PREDetector (Hiard *et al.* 2007). The set here represents the largest one which includes only sequences located in the appropriate position to obscure the -35 (or in one case the -10) site preceding coding sequences. A related sequence is found in the *txt* cluster as mentioned above (**5.2.2**) which diverges from this consensus.

Two of these target sites appear to encode uncharacterised proteins with similarity to known regulators. SCAB19291 is thought to encode a GntR family regulator. This family often has a C-terminal domain which binds a small molecule, along with the conserved N-terminal DNA-binding domain. SCAB24341 may encode a sucrolytic enzyme and is followed by two coding sequences probably for a sensor-regulator pair. These sequences with known regulatory function are potentially the most interesting in this group as they could function in a signalling cascade. In this organism, that could be a lead for investigating the triggers of pathogenicity. Such sequences might for example integrate signals indicating proximity of the plant host with the iron deficiency signal to regulate a further suite of appropriate genes.

The last two sites in Table 5-2 are of interest because they are associated with the coding sequences for the *nrps5* cluster, which may encode enzymes for biosynthesis of a peptide siderophore (see **6.2.5.3**). The major biosynthesis protein SCAB85471 in that cluster appears to have another closely related sequence to the IdeR target site found in *Mycobacterium tuberculosis* (Gold *et al.* 2001) in the preceding intergenic region. It may be that this is also bound and thus repressed by an IdeR-like protein, or perhaps that it diverges from the biological search pattern in the same way it diverges from the computational one.

The coding sequence most likely to function as the iron-dependent repressor is SCAB51401, homologues of which have been named *ideR* and *desR*. This is the most closely related predicted protein in the *S. scabies* 87.22 genome to diphtheria toxin repressor protein DtxR (Wennerhold and Bott 2006) and *Mycobacterium tuberculosis* iron-dependent repressor IdeR (Gold *et al.* 2001). The target sequence is not found upstream of the regulator's own sequence.

5.2.5 Complex products in pathogenicity

As might be expected in a streptomycete pathogen, biosynthesis of complex natural product appears to play a large part in pathogenicity in this organism. The coding sequences and products from these biosynthesis gene clusters are discussed in detail in Chapter 6.

The peptide siderophore pyochelin is a virulence factor in *Pseudomonas aeruginosa* infections, and an identical or very similar molecule has been identified in culture supernatant of *S. scabies* 87.22, and is likely to be produced by coding sequences SCAB1381-1481, as described in **6.2.4**. Another gene cluster has been identified (SCAB79601-79721 further described in **6.2.5.1**) which appears to be closely related and not identical to the one known to produce the coronafacic acid component of the *Pseudomonas syringae* phytotoxin coronatine, and this also may play a role in pathogenicity. A second peptide siderophore gene cluster has also been identified (**6.2.5.3**). A hybrid PKS/NRPS system found in RD35 may also be implicated in pathogenicity; none of these were previously known to be encoded by the genome of *S. scabies* 87.22.

Some of the complex products conserved in the non-pathogenic streptomycetes may also have a role in pathogenicity in *S. scabies* 87.22. Desferrioxamines are known to be involved in fireblight virulence in *Erwinia amylovora* (Dellagi *et al.* 1998; Expert 1999), and are thought to play a role in defence against the oxidative burst plant response to infection. Desferrioxamines may be expressed or regulated differently in *S. scabies* 87.22 from the non-pathogenic streptomycetes, and this may contribute to virulence. Concanamycin is also known to be produced by several *Streptomyces* strains, and given its specific inhibitory effect on V-type ATPases, it may also have a role in pathogenicity in *S. scabies* 87.22.

5.3 Conclusions

Thirty-five regions of difference (RD) greater than 10 kbp in length have been identified in the genome of *S. scabies* 87.22. Known virulence factors, thaxtomin biosynthesis (RD9) and the *nec1* locus and tomatinase (RD34) were identified in two of these regions. A further region (RD35) contains putative extracellular degradative

enzymes and coding sequences for biosynthesis of complex products which could have a role in pathogenicity.

Genetic information encoded in the pathogenicity island previously sequenced from *S. turgidiscabies* str Car8 appears in at least two locations in the genome of *S. scabies* 87.22. A mechanism is suggested for division of the PAI into two sections by homologous recombination beginning at identical repeats of 771 bp found at base positions 3744492..3745254 and 8894945..8895716, possibly during arm exchange.

Repeats often associated with RTX/haemolysin proteins were identified in the *S. scabies* 87.22 genome, but these appear unlikely to contribute to pathogenicity, due to the absence of the relevant N- and C-terminal domains in the structural elements, and the absence of activator and transport system proteins necessary for the function of this family of proteins. Pectin breakdown enzymes likewise are unlikely to contribute to virulence in this organism: there is no reported pectin breakdown phenotype, and the possible pectin breakdown enzymes found in *S. scabies* 87.22 do not appear to be significantly different to those identified in non-pathogenic streptomycetes *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680.

Operator sequences potentially bound by an iron-dependent repressor have been identified, as has the DtxR/IdeR homologue proposed to encode this repressor. Given the importance of the regulon activated by low iron conditions, coding sequences in this group may prove to be important in virulence. The hypothesis that iron-regulation is a factor in pathogenicity traits could be tested by investigating whether endogenous supply of iron switches off pathogenicity-related genes, and whether those identified as near to iron-regulated sequences are expressed under conditions of low iron availability. Since the regulatory mechanisms of streptomycetes are so intricate, with so many kinds of regulation, it is possible that iron availability is only one of several signals to be integrated for initiation of pathogenicity. This complexity in the signalling network complicates testing parts of the system but a good experimental design may indicate a role for these sequences.

Biosynthesis of complex natural products is known to be important for pathogenicity in this organism, as might be expected of a streptomycete pathogen. Sequences

implicated by conserved domains or similarity as involved in biosynthesis of complex natural products are discussed in detail in **Chapter 6**.

5.3.1 Method evaluation

5.3.1.1 Iron regulon prediction

In this work a brief study was undertaken beginning with patterns from previously-identified iron box sequences (Gunter *et al.* 1993; Wisedchaisri *et al.* 2004; Yellaboina *et al.* 2004a). Regulon prediction is an exciting field and a software tool such as PREDetector (Yellaboina *et al.* 2004b; Hiard *et al.* 2007) have been used with promising results (Rigali *et al.* 2004; Rigali *et al.* 2006; Hiard *et al.* 2007) and show potential for good use of genome sequence data for investigations into the biology of organisms.

6 Results – gene clusters for complex product biosynthesis

6.1 Introduction

This chapter describes gene clusters in *S. scabies* 87.22 identified as likely to encode enzymes for complex product biosynthesis, according to methods described in **Chapter 3, Methods for in-depth study of gene clusters**. It includes only those which are not found in either of *"S. coelicolor"* A3(2) or *S. avermitilis* MA-4680, according to tblastx comparison (using methods described in **3.1.1 Is this cluster the same as that cluster?**)

For each gene cluster identified as possibly encoding proteins involved in biosynthesis of a complex natural product, an 'arrow view' of the coding sequences is presented. This represents the estimated limits of the gene cluster, based on inference of co-transcription or co-regulation from the genome, as described in **3.1.4 Boundaries of gene clusters**. A module map has been drawn for multi-domain proteins such as those found in non-ribosomal peptide synthetase (NRPS) and polyketide synthase (PKS) systems, depicting domains identified by similarity and consensus model matches (For more detail see **3.2.1 Module map**). Where possible, a drawing of any predicted product molecule or biosynthetic pathway is included.

Six gene clusters are described in detail. Two of these are treated as proven functional, from demonstrations of previous researchers of the complex products produced: those encoding the enzymes for thaxtomins (**6.2.3.1)**, and those for concanamycins (**6.2.3.2**). One other has some evidence supporting the predicted product, a pyochelin-like molecule (**6.2.4)**. Another has sufficient similarity to a known cluster to predict something about the products likely to be encoded (**6.2.5.1**). Two more clusters contain NRPS multienzymes (**6.2.5.2** and **6.2.5.3**), for which some predictions may be possible about the substrates and activity of enzymatic domains. Seven further clusters have not been studied in detail, but any striking features are summarized below (**6.2.2)**.

6.2 Results and discussion

6.2.1 Method development

Adenylation domains activating other substrates

Several adenylation domains were identified within the genome of *Streptomyces scabies* 87.22 with above-threshold matches to PF00501 and not showing the characteristic motifs of either the amino acid or aryl acid activating clades. To test the phylogenetic grouping of these domains a set of sequences was constructed including the two reference domains 1AMU (for amino acids) and 1MD9 (for aryl acids), and five further sequences. Four domains with characterised adenylation function on other substrates were retrieved from INSDC: feruloyl-CoA synthetase of *Delftia acidovorans*, and O-succinylbenzoic acid-CoA ligase of *Mycobacterium tuberculosis*, and acetoacetyl-CoA syntetase from *Sinorhizobium meliloti*, and the adenylation domain proposed to activate the nitro-aryl acid in the aureothin biosynthesis cluster. The sequence of the first enzyme in the adenylate-forming family to be crystallised - *Photinus pyralis* luciferase (Conti *et al.* 1996) - has also been included.

Conserved sequence motifs in alignments match the groups shown in Figure 6-1. Sequence SCAB72991, SCAB0751 and the second adenylation domain in SCAB43961 all have the expected motifs for activating an amino acid (Motifs A3, A4, A5 in Figure **3-3**) SCAB63261 also has several of the expected motifs for activating an amino acid, but the next sequence moving out from that branch in Figure 3-7, SCAB1531, does not have the conserved motifs. So SCAB63261 and all the sequences towards the GrsA end of the tree (Figure 3-7) have been assumed to activate an amino acid.

The aryl-acid activating clade which includes the reference sequence DhbE (PDB: 1md9) appears to be much smaller (Figure 3-7). Only SCAB1411 and SCAB1671 show the (DhbE-331)**GMAEG**(DhbE-335) motif expected. Note that SCAB51551 and SCAB79361 have a somewhat similar primary sequence at this position.

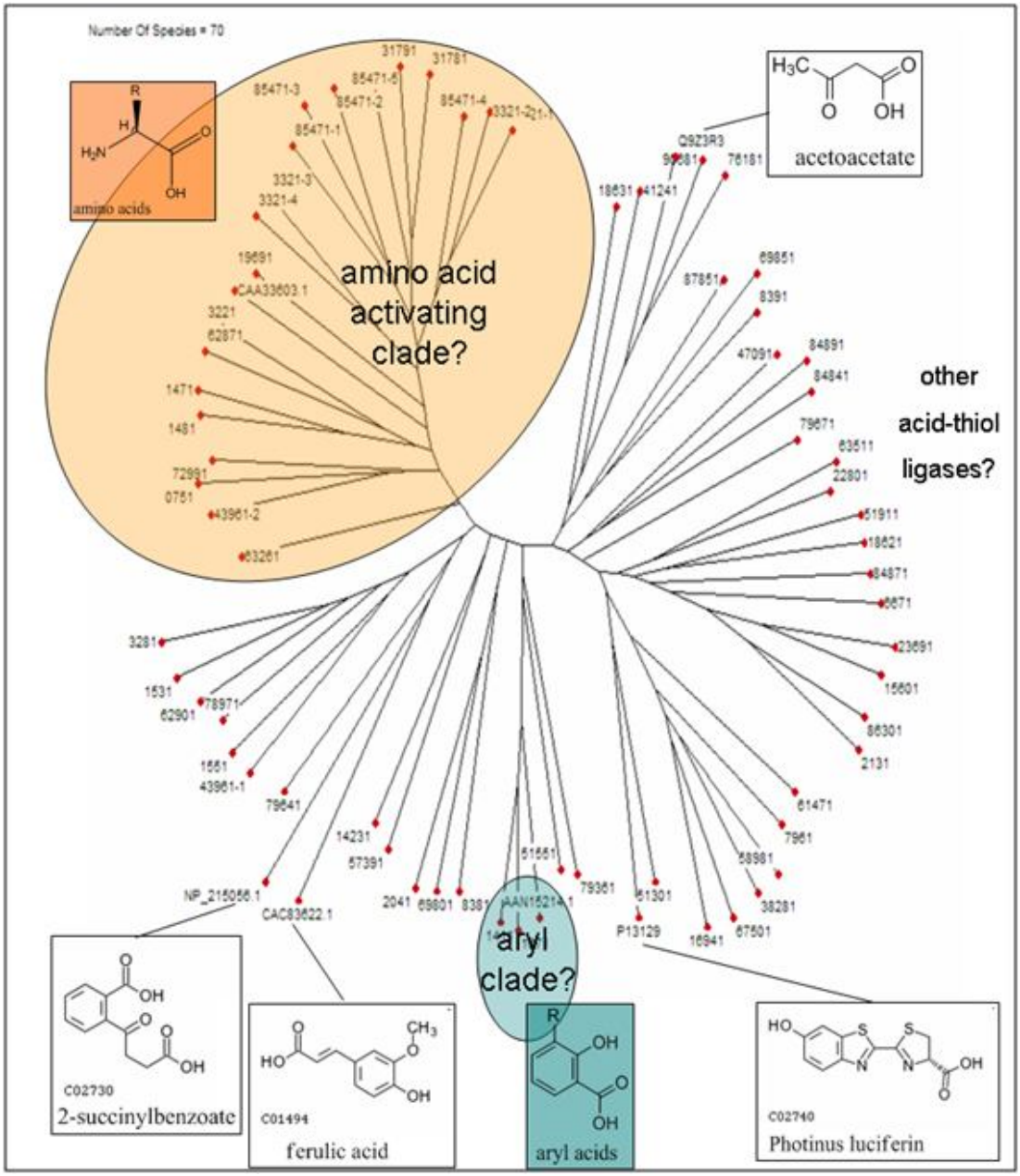


Figure 6-1 Adenylation domains from S. scabies 87.22 matching PF00501 with additional sequences for comparison. CAA33603=1AMU; AAN15214.1=1MD9; S. scabies domains appear by digits alone so SCAB1551=1551 on the diagram.

There appears to be a clade amongst the *S. scabies* 87.22 sequences which includes both NP_215056.1 and CAC83622.1 (Figure 3-7). It is not clear what the significance of this clade is so it has not been used for diagnostic alignment. The photinus-luciferin-binding domain (P13129, Figure 3-7) is from a eukaryotic protein, unlike all the rest which are

from eubacteria and was intended as an out-group, although one of the *S scabies* domain (51301, Figure 3-7), SCAB51301 appears most closely related to it. It is not

clear what the significance of this is, if any, perhaps function-driven convergent evolution or horizontal transfer.

Several sequences from *S. scabies* 87.22 group with the acetyl-coA synthetase sequence from *Sinrhizobium meliloti* (Q9Z3R3, Figure 3-7) and these might be assumed to perform similar functions in core metabolism. This clade of sequences in *Streptomyces scabies* 87.22 would more accurately resolved by comparison with a much larger number of sequences from other characterised acid-thiol ligase enzymes, but has been used in this work as a quick way to check whether a sequence falls into either the amino or aryl acid activating domains.

Peptidyl carrier protein (thiolation) domains

Carrier domains have been observed to have a conserved sequence motif indicating the enantiomer of the amino acid they carry (Linne *et al.* 2001). Domains supporting incorporation of the D amino acid enantiomer were identified to have a conserved **GGDSI** motif in TycA (tyrocidine synthase (INSDB: P09095) of *Brevibacillus brevis* (Linne *et al.* 2001)). The conserved aspartate residue has been shown by mutational analysis to have a functional role in binding the D conformer; carrier domains involved in incorporating the L amino acid, the motif at the same position in TycA (INSDB: P09095) is **GGHSL** (Linne *et al.* 2001).

In this work a similar pattern was found when comparing carrier domains likely to be carrying L and D enantiomers of amino acids. Residues around the position aligning with Ser45 (PDB: 1DNY) in a multiple sequence alignment were examined. Domains where the motif has the form **GxHSx** have been predicted to carry L-amino acids and domains where the motif is **GxDSx** predicted to carry the D-amino acid. The fifth position in the motif (aligning with residue 46 (PDB: 1DNY) has been found to vary, but is consistently hydrophobic (Ile, Leu, or Val). This more degenerate conserved motif is probably a result of the evolutionary separation between *Brevibacillus parabrevis* and *Streptomyces scabies*; the second position in the motif is not conserved and the fifth position contains variation within the range of small and medium-sized hydrophobic residues.

Where domains matching PF00550 have been found in clusters studied in detail with divergent sequences not conforming to either of these reference has been made to

this in the text. (The carried amino acid could for example be glycine which does not have a stereocentre at the α carbon, due to having two identical groups).

Condensation domains

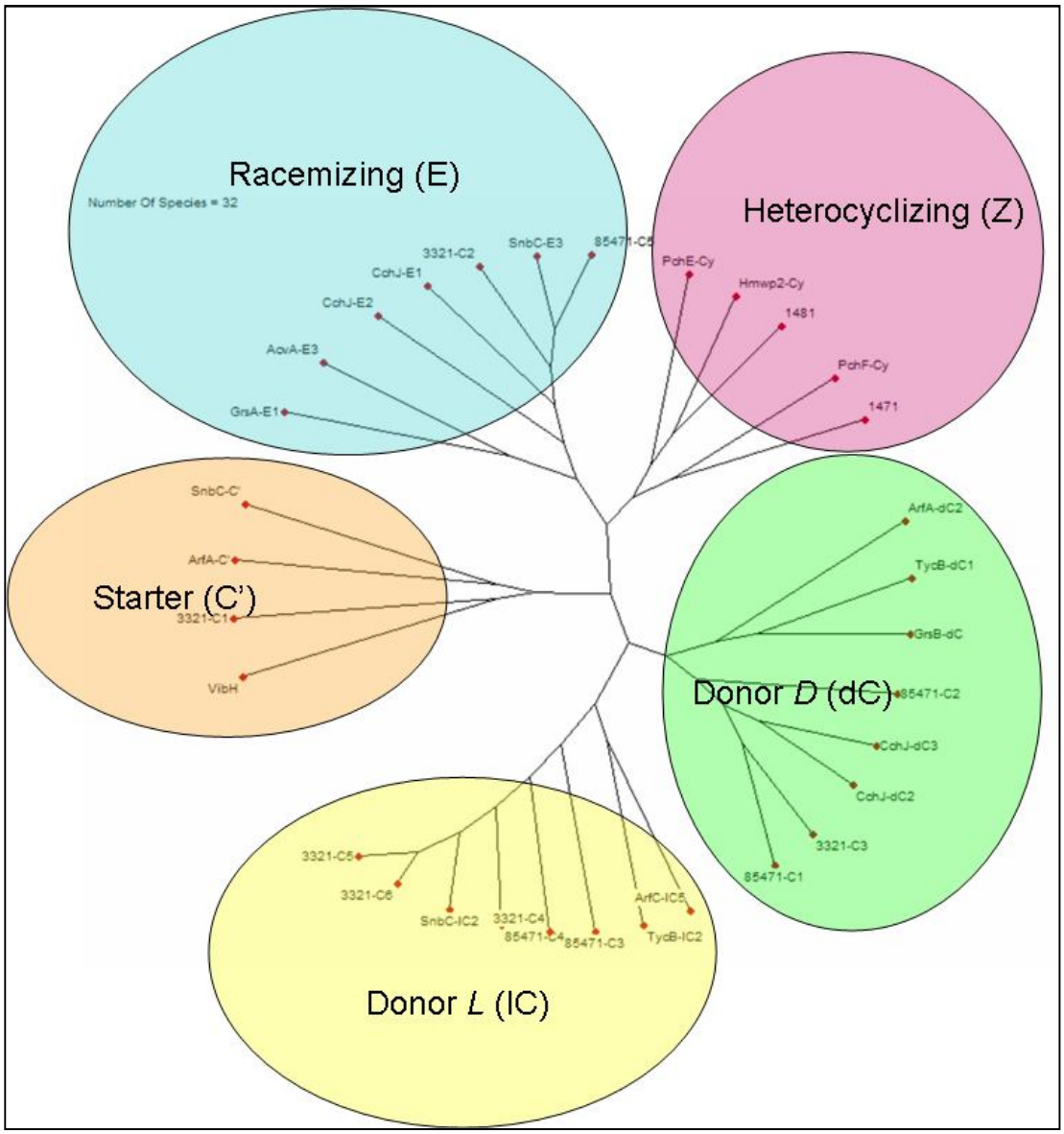


Figure 6-2 Condensation domains matching PF00668 from *S. scabies*. Reference domains from INSDC to illustrate subfamilies. Stepwise (radial) tree with leaf nodes extended for clarity.

PF00668 domains fall into five almost entirely monophyletic groups (Rausch *et al.* 2007), the exception being a few apparent instances of convergent evolution where domains in the dC clade appear to have reverted to select L amino acids: Starter (C') domains condense a fatty acid or other residue onto the start of the chain; condensation reactions with a D-amino acid donor position (dC); and those with an L-amino acid donor (lC) are similar to each other; domains catalyzing

heterocyclization of the incorporated residue (Z); and so-called 'epimerase' domains which racemize the next incorporated amino acid (E) also. These clades are all represented in the genome of *S. scabies* 87.22 (Figure 6-2). Domains matching the Pfam model PF00668 have thus been tested via a multiple sequence alignment to predict which condensation subfamily they fit into. It is necessary to take into account all data about which reaction is catalyzed, because of the potential for convergent evolution as mentioned above.

Where a condensation domain appears first in an NRPS multienzyme, the first element in the chain may be a fatty acid (Roongsawang *et al.* 2003). This prediction is supported if an appropriate adenylation domain is found adjacent and if the condensation domain groups with the Starter or C' clade predicted to condense β-hydroxy-carboxylic acids (Rausch *et al.* 2007). As VibH is unusual, catalyzing condensation of 2,3-dihydroxybenzoic acid to the primary amine of norspermidine (Keating *et al.* 2002) its position in the starter C' clade (Figure 6-2) is not unexpected.

6.2.2 Summary of 14 nonconserved complex product gene clusters

Of the total 29 gene clusters identified in this study as likely to encode proteins involved in complex product biosynthesis, nearly half (n=14) are not found in either "*S. coelicolor*" A3(2) or in *S. avermitilis* MA-4680. These 14 clusters are summarized in this chapter (Table 6-1). Comparisons of whole genome alignments between the three genomes (for example Figure 4-4) showed several insertion or deletion gaps in synteny, in which the non-conserved gene clusters are found.

In one of these regions, perhaps an insertion in *S. scabies* 87.22 by a mobile pathogenicity island (see Introduction **1.3.6**), the thaxtomin biosynthesis genes are found as described in **5.2.2**. A conservon is also found at this locus alongside the thaxtomin biosynthesis genes, and a cluster which might encode a LanB/C type lantibiotic, SCAB31961-SCAB32051. This region falls within the parallel insertion point SCAB31731-SCAB33431: in alignment of the three genomes "*S. coelicolor*" A3(2) has actinorhodin biosynthesis at this point, and *S. avermitilis* MA-4680 has another NRPS system. Future investigations might be able to determine whether there are any special properties of this locus, such as might allow uptake or recombination at a higher frequency. A hybrid NRPS/PKS cluster was identified and

is briefly described (**6.2.4.4)** which is found in RD35, a region of difference identified in comparative studies between the three available genomes.

Biosynthetic system	Metabolite	Coding sequences in *S. scabies 87.22* (systematic identifiers)	Gene cluster size estimate /kb
NRPS	pyochelin-like siderophore?	SCAB1381-SCAB1571	35
NRPS	novel lipopeptide?	SCAB3281-SCAB3361	34
Class II DAHP synthase	2-amino-3-hydroxybenzoic acid?	SCAB12021-SCAB12111	12
Mixed	unknown	SCAB19681-SCAB19731	8
NRPS	thaxtomins	SCAB31761-31841	19
Lantibiotic	novel lanBC-type lantibiotic	SCAB31961-SCAB32051	14
Hybrid NRPS/PKS system	unknown	SCAB43961	13
Hybrid NRPS/PKS system	unknown	SCAB62901-SCAB63011	14
Mixed	unknown	SCAB63251-SCAB63401	8
Mixed	unknown	SCAB69772-SCAB69871	13
Hybrid NRPS/PKS system	unknown	SCAB78961-SCAB78981	8
Mixed type I/II PKS	coronafacic acid derivative?	SCAB79591-SCAB79721	26
Type I PKS	concanamycins	SCAB83871-SCAB84091	95
NRPS	peptide siderophore?	SCAB85461-SCAB85521	42

Table 6-1 Summary of gene clusters in S. scabies 87.22 identified in this work as not conserved in "S. coelicolor" A3(2) and S. avermitilis MA-4680. NRPS=nonribosomal peptide synthetase; PKS=polyketide synthase.

In one of these regions, perhaps an insertion in *S. scabies* 87.22 by a mobile pathogenicity island (see Introduction **1.3.6**), the thaxtomin biosynthesis genes are found as described in **5.2.2**. A conservon is also found at this locus alongside the thaxtomin biosynthesis genes, and a cluster which might encode a LanB/C type lantibiotic, SCAB31961-SCAB32051. This region falls within the parallel insertion point SCAB31731-SCAB33431: in alignment of the three genomes "*S. coelicolor*" A3(2) has actinorhodin biosynthesis at this point, and *S. avermitilis* MA-4680 has another NRPS system. Future investigations might be able to determine whether there are any special properties of this locus, such as might allow uptake or recombination at a higher frequency. A hybrid NRPS/PKS cluster was identified and is briefly described (**6.2.4.4)** which is found in RD35, a region of difference identified in comparative studies between the three available genomes.

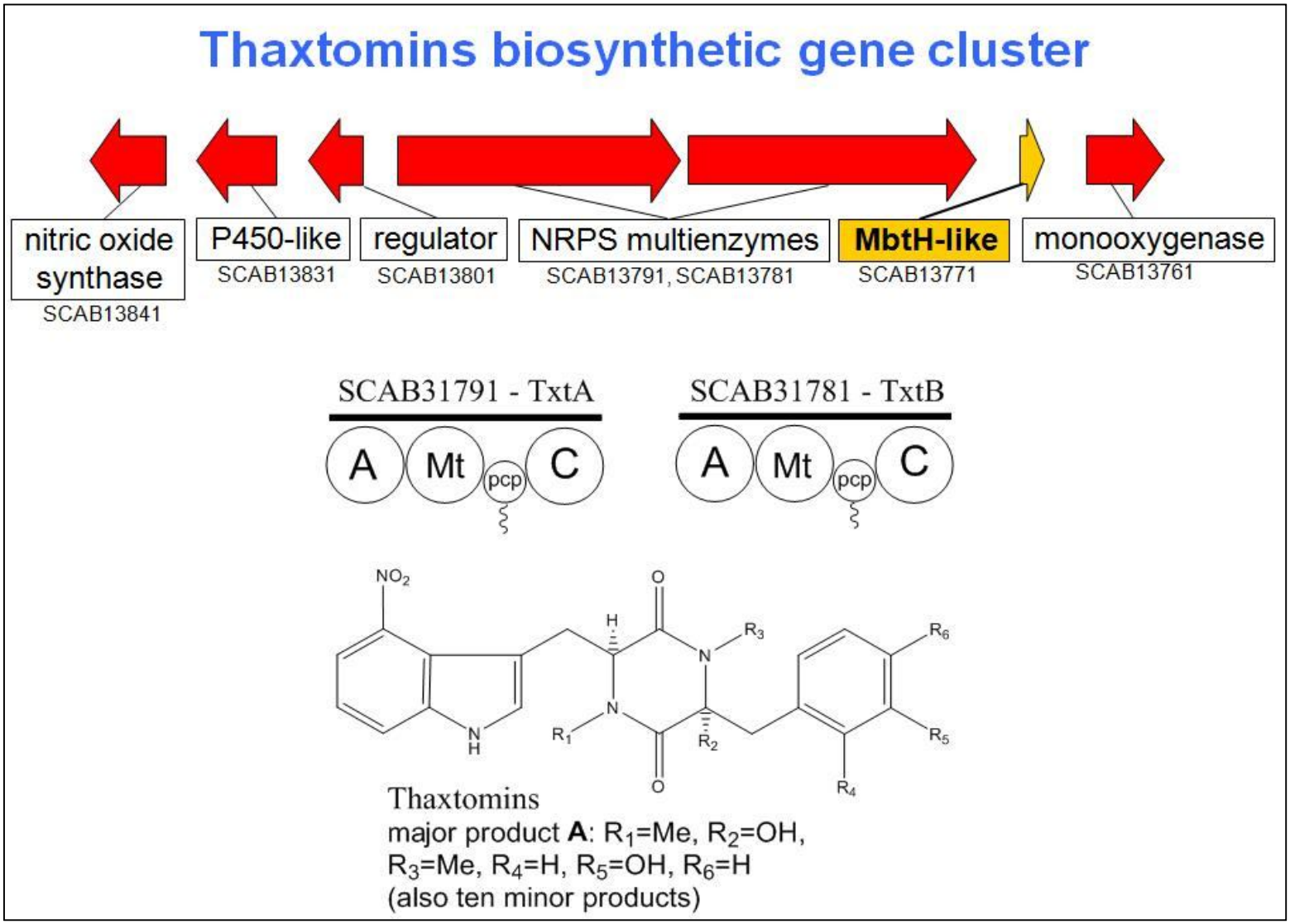


Figure 6-3 Summary diagram showing thaxtomins biosynthesis genes in *S. scabies* 87.22, architecture of multienzyme proteins, and structure of major product thaxtomin A.

6.2.3 Gene clusters with known products: thaxtomins, concanamycins

S. scabies is known to produce dipeptide phytotoxin thaxtomins (Loria *et al.* 1995), and plecomacrolide concanamycins A and B (Natsume *et al.* 2001). Neither of these groups of products has been shown to be produced by "*S. coelicolor*" A3(2) or *S. avermitilis* MA-4680, nor have the genes likely to encode the enzymes for the production pathways been discovered in either of those genomes.

6.2.3.1 Thaxtomins SCAB31761-SCAB31841

Gene clusters for biosynthesis of thaxtomins have been sequenced from scab-causing organisms including *S. acidiscabies* 84-104 (Healy *et al.* 2000), and *S. turgidiscabies* Car8 (Kers *et al.* 2004). Genes for thaxtomin biosynthesis were identified as part of the mobile pathogenicity island sequenced from *S. turgidiscabies* Car8 (Kers *et al.* 2005). The biosynthesis gene cluster for thaxtomins (6 1) was identifiable during annotation by comparison with known sequences for biosynthesis of these products (INSDC: AAG27088; AY707081). The expected enzymes all appear to be present, and encoded in the same arrangement. One additional coding sequence has been discovered, *txtH*. This small coding sequence, SCAB13771, is found just 3' of genes for the two NRPS multienzyme *txtAB*.

MbtH-like protein TxtH

This extra coding sequence *txtH* in *S. scabies* 87.22 (Table 6-2) was not annotated in previous sequences for biosynthesis of thaxtomins from the *S. turgidiscabies* strain Car8 pathogenicity island AY707081 (Kers *et al.* 2005). Frame plot and correlation scores indicated a coding sequence between *txtB* SCAB31781 and *txtC* SCAB31761. This appears to carry a strong match to Pfam model PF03621 'MbtH-like protein'. The gene appears to be present on examination of the related sequence from *S. turgidiscabies* strain Car8.

MbtH protein was first identified in the biosynthesis cluster for siderophore mycobactin produced by *Mycobacterium tuberculosis* (Quadri *et al.* 1998). This type of coding sequence appears to be widespread in NRPS gene clusters (Stegmann *et al.* 2006). Deletion of the MbtH-like gene from the gene cluster for production of

glycopeptide balhimycin did not affect production (Stegmann *et al.* 2006); but another MbtH-like gene PA2412 is necessary

for the production or secretion of pyoverdine at normal levels (Drake *et al.* 2007). A PA2412-deletion mutant was able to use an exogenous supply of pyoverdine, indicating that without PA2412 the mutant was still able to take up and utilize the iron-pyoverdine complex (Drake *et al.* 2007).

Predicted protein name	*S. scabies* coding sequence	*S. turgidiscabies* pathogenicity island coding sequence	% amino acids identical	function
TxtC	SCAB31761	stPAI031	90	P450 family monooxygenase
TxtH	**SCAB31771**	**-**	**79***	**MbtH-like?**
TxtB	SCAB31781	stPAI030	89	NRPS
TxtA	SCAB31791	stPAI029	90	NRPS
TxtR	SCAB31801	stPAI027	78	cellobiose-dependent regulato
-	SCAB31831	stPAI022	87	P450-like?
Nos	SCAB31841	stPAI021	100	nitric oxide synthase

Table 6-2 Comparison of proteins encoded in thaxtomins/NO gene cluster in *S. scabies* 87.22. *no coding sequence defined in *S. turgidiscabies* PAI annotation; figure derived from comparison with translation from *S. turgidiscabies* Car8 PAI nucleotide sequence.

Further deletion studies have shown that only one, any, of several MbtH-like proteins in a genome need be functional for biosynthetic activity to occur in all the MbtH-containing clusters (Wolpert *et al.* 2007). This suggests that MbtH-like proteins may be a mechanism for cross talk between biosynthetic pathways (Lautru *et al.* 2007), perhaps as a mechanism for co-ordinating biosynthetic activities. This possible co-ordination function makes the discovery of an MbtH-like protein in the thaxtomin/NO biosynthesis cluster interesting; this small protein could play a role in co-ordinating pathogenicity traits.

Insertions/deletions within txt cluster

Comparison by tblastx and blastn between *S. scabies* 87.22 and related sequence from *S. turgidiscabies* Car8 pathogenicity island revealed several indels around the biosynthetic gene cluster for thaxtomins (Figure 6-2). These regions appear to contain pseudogenes and gene fragments. Between the *txt* and *nos* coding sequences, several DNA sequence matches generate high scoring pairs in blastn comparison. These (example score 155, 93% nucleotide identity) may indicate selection for function, such as a binding site.

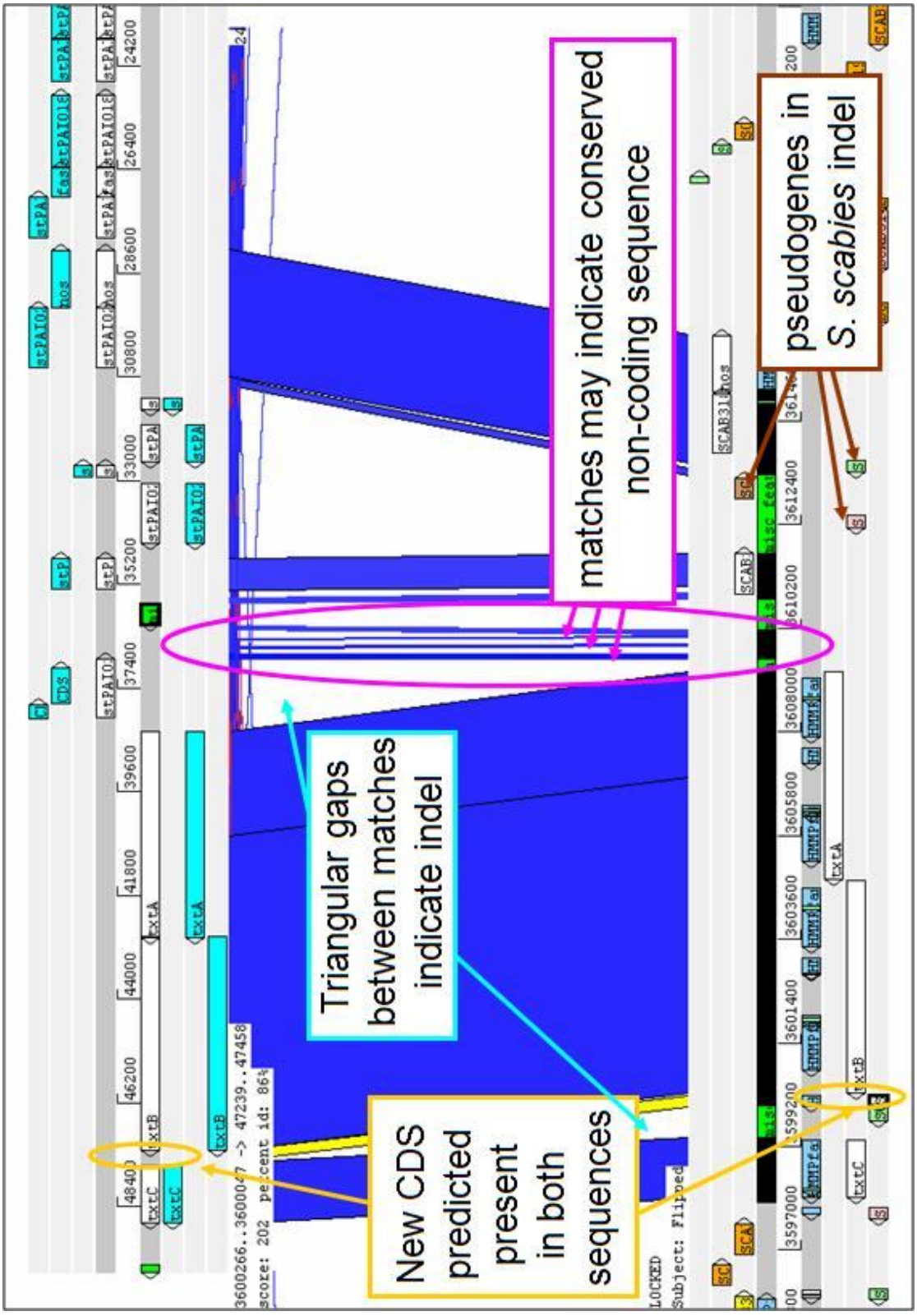


Figure 6-4 Comparison with blastn between *S. turgidiscabies* Car8 pathogenicity island (top) and related region of *S. scabies* 87.22 genome (bottom), visualized in ACT. Several indels are apparent from triangular gaps in alignment. Coding sequences with gold-edged arrows indicate position of MbtH homologues. Cyan and magenta features relate to indels between the two sequences; brown features indicate pseudogenes found near of the in/dels.

The differences between the coding sequences for the *txt* cluster genes are interesting because they contrast with the very high level of sequence conservation between sequences around the *nec1* locus which appear to be transferred between strains with high fidelity (Bukhalid *et al.* 2002). The apparent divergence between the two sequences for biosynthesis of thaxtomins implies that the lineages have been separated for long enough to accrue differences which may be neutral in selective effect. It would be interesting to find out whether the high scoring alignment between the two nucleotide sequences has any obvious significance, for example by examining the pathogenicity phenotype of a mutant with this stretch of DNA removed.

6.2.3.2 Concanamycins SCAB83841-SCAB84141

A gene cluster for biosynthesis of concanamycin A has been sequenced in *S. neyagawaensis* ATCC 27449 (Haydock *et al.* 2005). That sequence has been used in this work for comparison with the cluster in *S. scabies* 87.22, which is proposed to encode enzymes for biosynthesis of concanamycins A and B. *S. scabies* 87.22 and related strains are known to produce concanamycins A and B (Natsume *et al.* 2001), but genes have not previously been identified from *S. scabies*. All the coding sequences identified in biosynthesis in the *S. neyagawaensis* cluster are present, although some are reversed in organisation, as described below.

An AfsR-family regulator protein sequence associated with this cluster contains a TTA codon. It has been suggested that availability of the rare tRNA with the anticodon matching TTA is part of a mechanism which could co-ordinate the biosynthesis of many complex products with morphological development in *"S. coelicolor"* (Hesketh *et al.* 2007), possibly also in other organisms.

Global alignments of protein sequences for PKS multienzymes in this cluster showed high levels of similarity between *S. scabies* 87.22 and those in *S. neyagawensis*. An apparent gap in alignment between SCAB83901 and *conF* is visible on the tblastx comparison (Figure 6-5), but this is accounted for by two short (2 and 3 amino acid) gaps in the alignment within a few hundred residues. These are not located in regions thought critical for function. The score of the alignment for that region of the protein drops below the score >600 threshold used for display.

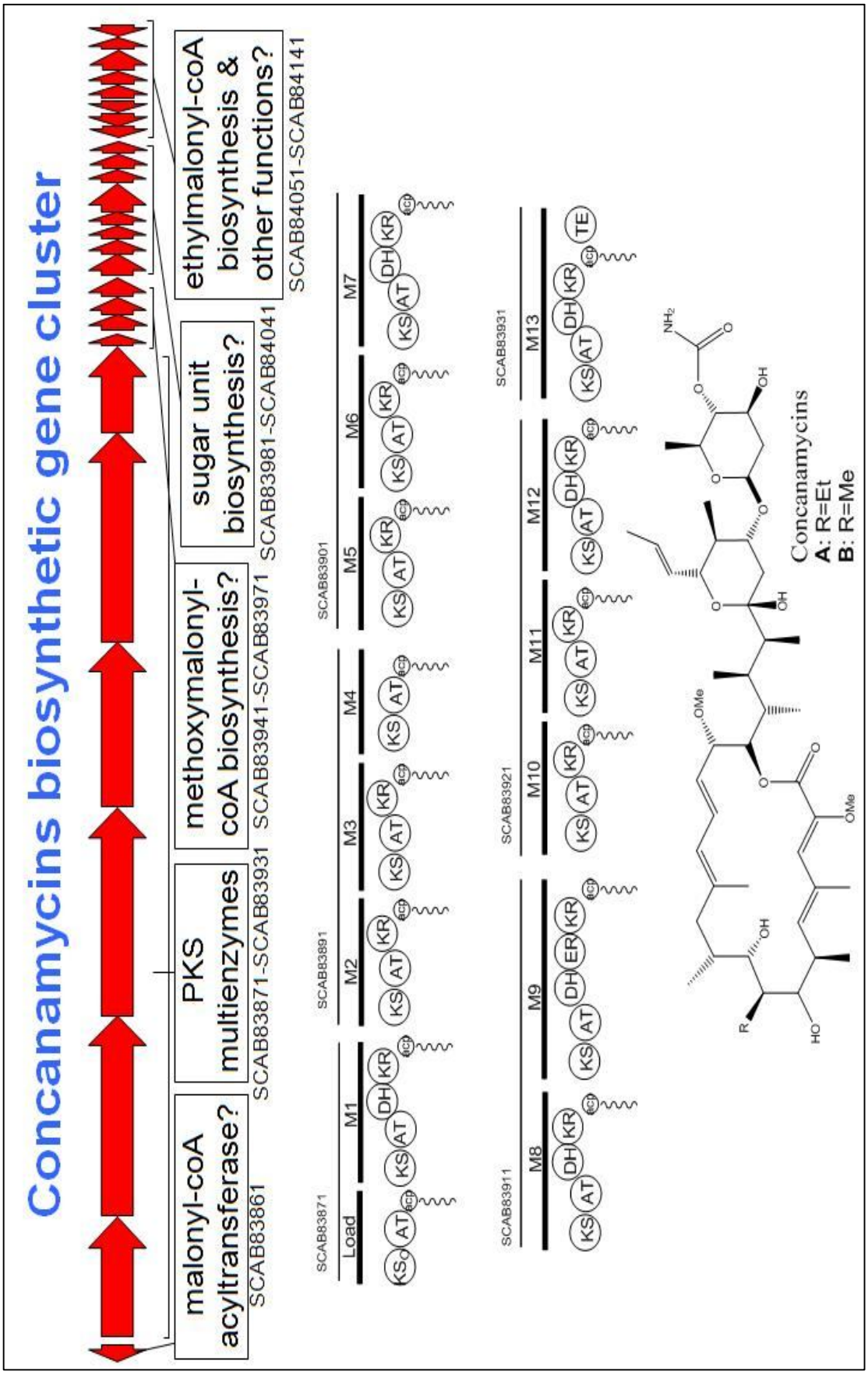


Figure 6-5 Summary of concanamycins biosynthetic gene cluster in *S. scabies* 87.22.

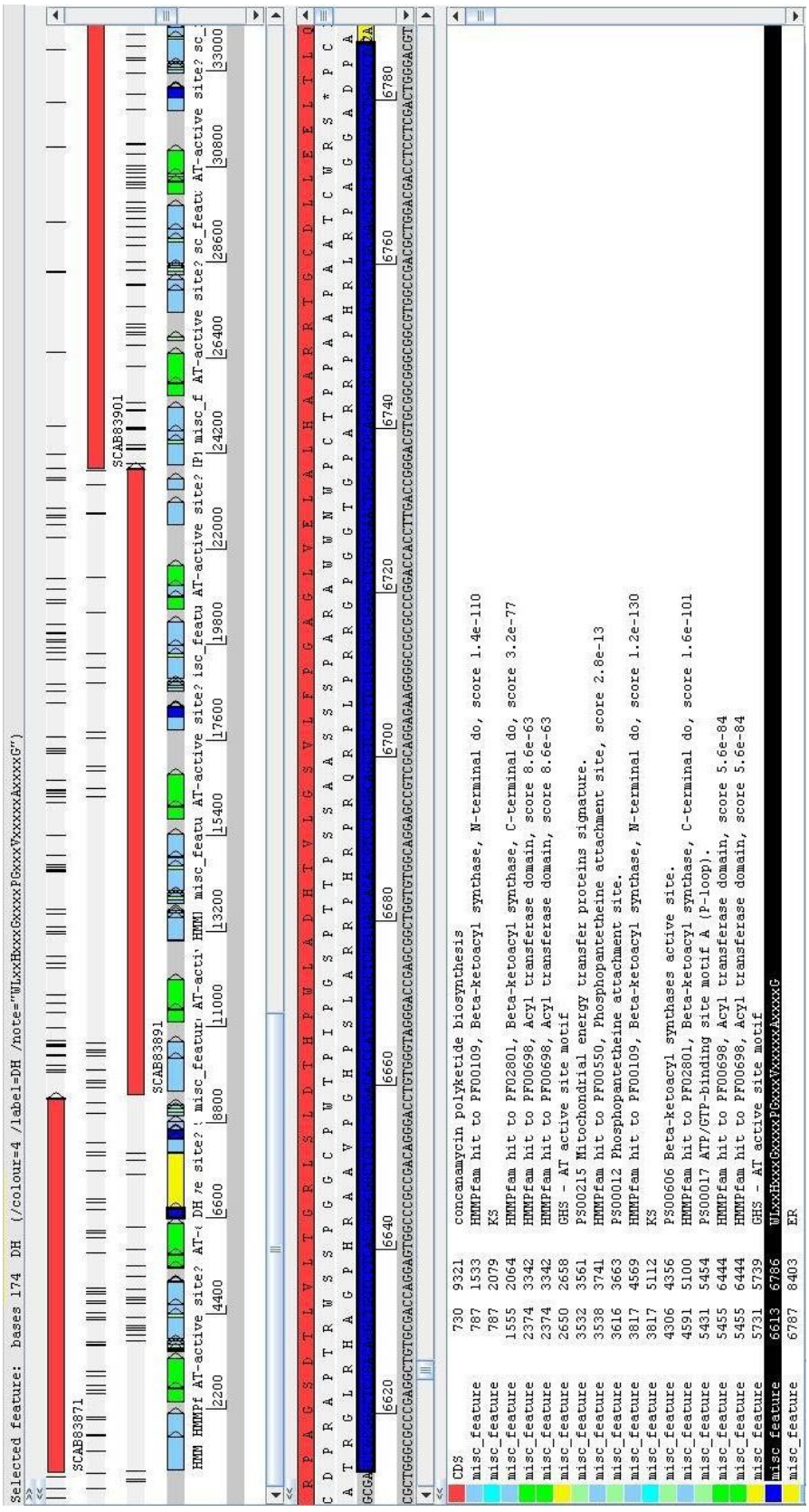


Figure 6-6 Illustration of in-depth study of predicted domains in SCAB83871 showing annotation of predicted active sites. The selected feature, described in the top line of text, bold outline in the display window and highlighted in black in the feature list, is the pattern used to search for characteristic amino acid motifs typical of dehydrogenase domains.

Domain functions predicted identical

Close study of the modular architecture and domains within the six PKS multienzymes SCAB83871-SCAB83931 shows the same domains present in both clusters. These domains have been illustrated with coloured features in the file provided in the electronic appendices for illustration, `concanamycin_cluster.tab` (the cognate `.dna` file will be necessary to visualize these data in Artemis (Rutherford *et al.* 2000).) Comparison of the putative active site residues in the predicted domains (illustrated in Figure 6-6) with those depicted in the published analysis of the concanamycin A biosynthetic gene cluster in *Streptomyces neyagawaensis* ATCC 27449 (Haydock *et al.* 2005) confirms that activity or inactivity of the domains is also predicted to be identical.

It can be observed that the leucine residue in the KR active site previously supposed invariant (Wu *et al.* 2005) can apparently be substituted by valine without loss of function, given that *S. neyagawaensis* ATCC 27449 concanamycin PKS modules 2, 5, and 10 seem to have KR function and that production of the concanamycin products has been demonstrated in the past from both *S. neyagawaensis* and *S. scabies.*

Sugar unit biosynthesis genes inverted

From tblastx comparison (Figure 6-5), seven coding sequences in the cluster have been inverted between the *S. scabies* 87.22 and *S. neyagawaensis* lineages. These seven coding sequences are thought to encode enzymes for sugar unit biosynthesis (Haydock *et al.* 2005). Inversion might have inactivated these genes by disrupting regulatory arrangements, but proteomics data indicates that the inverted genes are expressed (R. Loria pers.comm.).

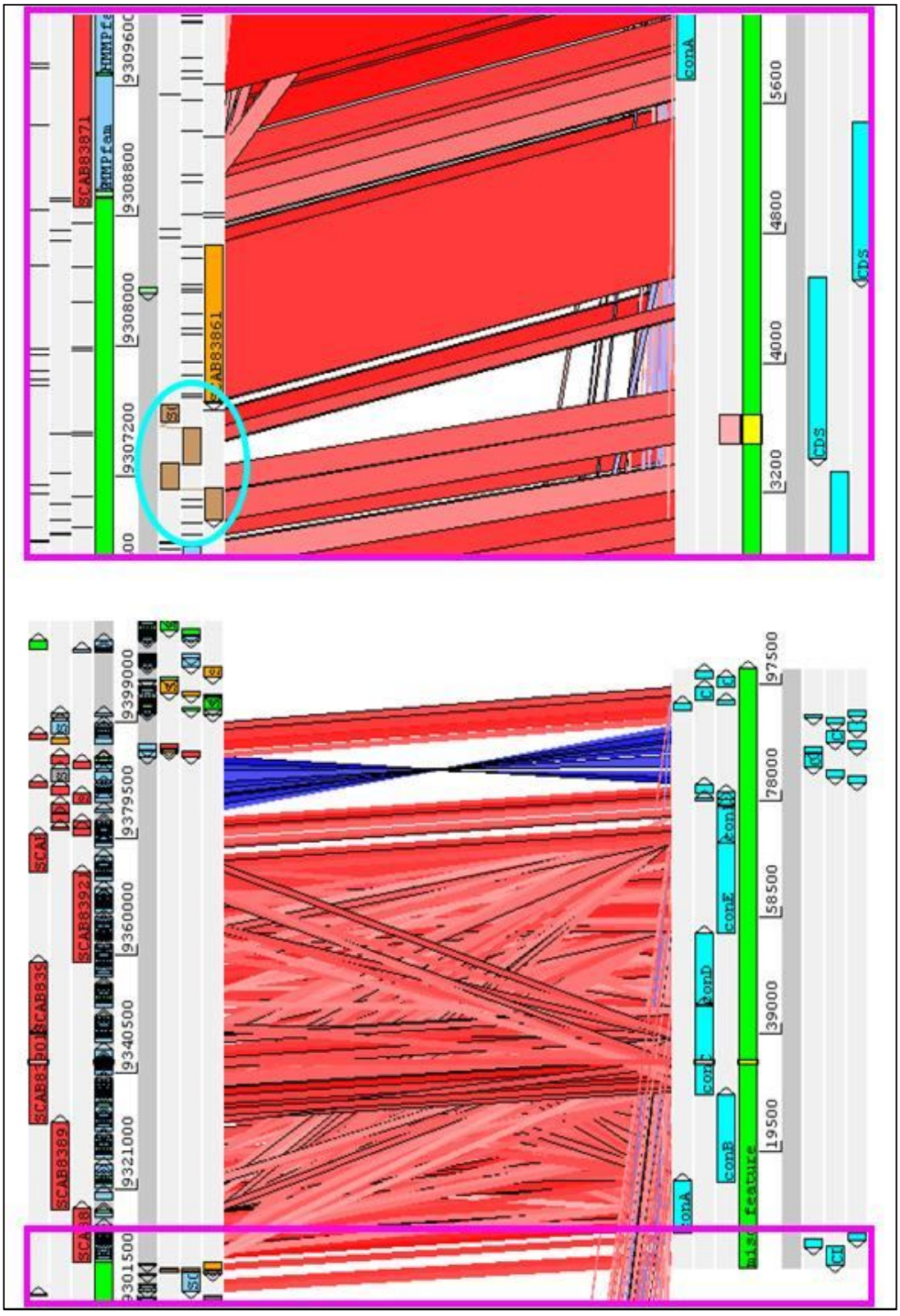


Figure 6-7 Comparison of biosynthesis gene clusters for concanamycins. (Left) Overview of comparison showing matches scoring >600 between *S. scabies* 87.22 cluster (top) and *S. neyagawaensis* concanamycin A cluster: inverted section indicated with dark blue rather than red stripes. (Right) closeup boxed region at left hand end of comparison showing pseudogene (cyan oval) in *S. scabies* 87-22.

This inversion of a section of DNA is an example of natural rearrangement of coding sequences, and could be involved in generating novelties upon which natural selection may be able to act. Such rearrangements could change function by interrupting coding sequences or their regulatory sequences, which could have selective advantages or disadvantages. Sequencing this region in several other organisms in both *S. scabies* and *S. neyagawaensis* species and other organisms from this region of the genus would allow discoveries about how widespread each arrangement is; examination of the products of the organism would indicate whether the rearrangement has any significance for function. The evolutionary processes by which natural product biosynthesis pathways are altered and rearranged is of interest because it has the potential for illuminating the origins and natural diversity of these remarkable substances.

6.2.4 Some evidence for production: pyochelin-like? SCAB1381-1481

A gene clusters found in the genome of *S. scabies* 87.22 has similarity to a sequenced gene cluster (Reimmann *et al.* 2001)known to produce peptide siderophore pyochelin. Two NRPS multienzyme proteins were predicted, apparently similar to those involved in pyochelin biosynthesis. Peptide siderophore pyochelin is known to be a virulence factor in *Pseudomonas aeruginosa* infections of vulnerable humans (Wang, J. *et al.* 1996), first isolated from *P. aeruginosa* PA01. [A mixture of 4 '*R*, 2 "*R*, 4 "*R* (pyochelin I) and 4 '*R*, 2 "*S*, 4 "*R* (pyochelin II) is produced]. Enantio-pyochelin is the optical antipode of pyochelin and has also been identified as a natural product (Youard *et al.* 2007) from *Pseudomonas fluorescens* CHA0.

In *P. aeruginosa* pyochelin is assembled from salicylic acid and two molecules of cysteine, the reaction catalyzed by NRPS multienzymes encoded by *pchEF* (Reimmann *et al.* 1998). PchABCD are responsible for production and activation of salicylate (Serino *et al.* 1995); PchG has been shown to have NADPH-dependent reductase function (Patel and Walsh 2001) and appears to be required for production; PchHI are thought to have transmembrane domains and show similarity to ABC transport system components (Reimmann *et al.* 2001).

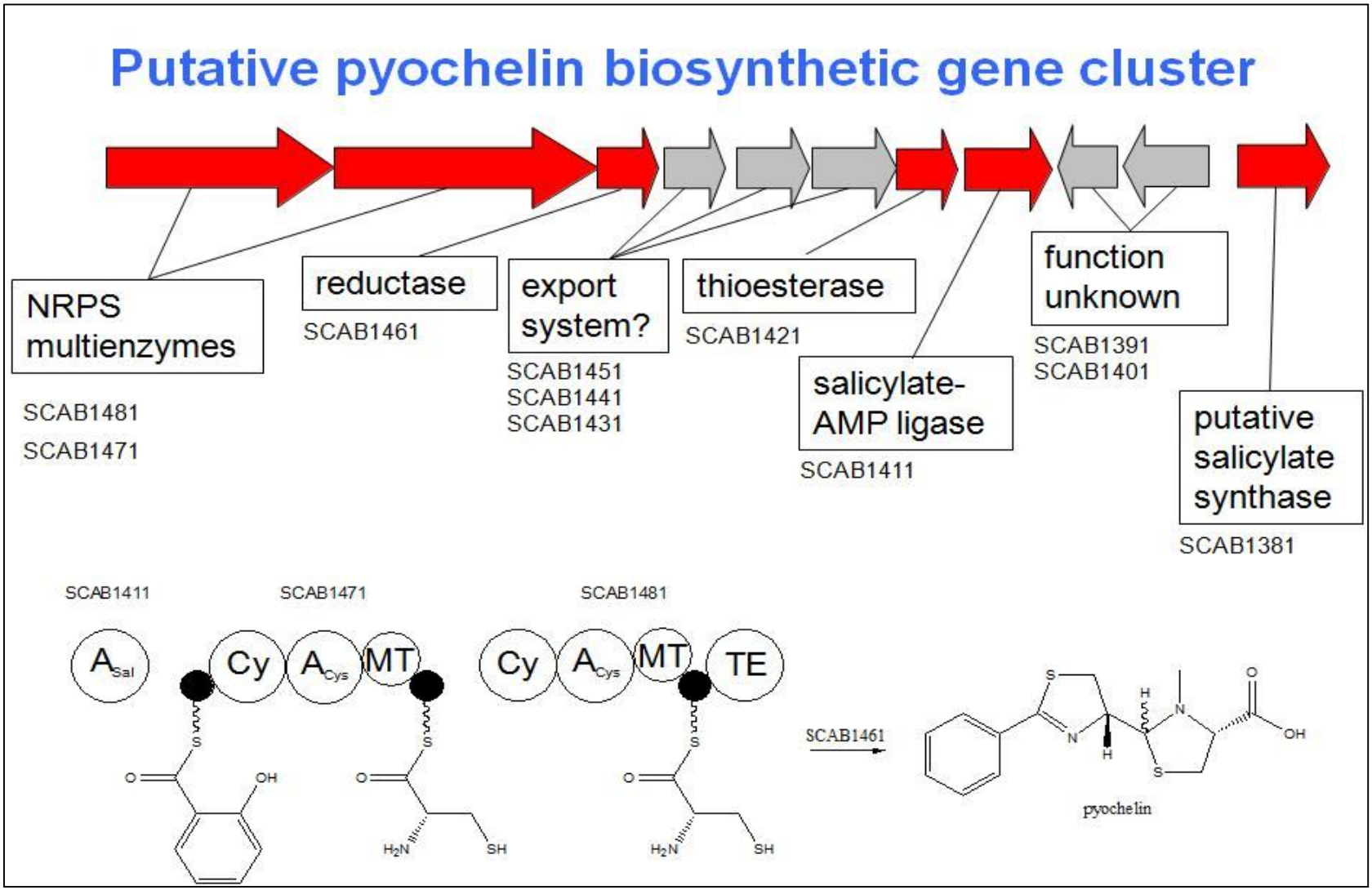


Figure 6-8 Summary of gene cluster possibly encoding enzymes for biosynthesis of pyochelin or a related product *in S. scabies* 87.22

Comparison with Pseudomonad pyochelin biosynthesis cluster

The gene cluster containing *pch*-like sequences appears to be 35kbp in size, identified in *S. scabies* 87.22 in the left chromosomal arm; limits of this gene cluster are proposed using tblastx comparison between *Pseudomonas aeruginosa* PA01 and *S. scabies* 87.22. A further 10-15 CDS upstream of this cluster include SCAB1531 predicted to encode a protein with adenylation (A domain) function, and it is possible that these may also be involved. The primary sequence of the protein SCAB1531 appears to fall outside the amino acid binding clade, and not within the aryl acid binding clade, and hence it is difficult to predict what kind of substrate is likely to be activated by it.

There are several differences between the two clusters (Figure 6-6), but the architecture of the synthetase multienzymes seems to be the same, with the exception of a region at the C-terminal of SCAB1481 and PchE.

CDS in *S. scabies* 87.22	predicted function	similar protein	% amino acids identical (fasta)
SCAB1481	NRPS multienzyme: C A Mt pcp (TE)	*Streptomyces coelicolor* A3(2), SCO7682	48.10
SCAB1471	NRPS multienzyme: pcp C A Mt pcp	*Streptomyces coelicolor* A3(2), SCO7683	46.29
SCAB1461	oxidoreductase	*Streptomyces coelicolor* A3(2), SCO7685	47.76
SCAB1451	transport system transmembrane component	*Frankia alni* ACN14a, unchar. transporter	46.50
SCAB1441	ABC transport system ATP-binding component	*Streptomyces coelicolor* A3(2), SCO7689	50.00
SCAB1431	ABC transport system ATP-binding component	*Streptomyces coelicolor* A3(2), SCO7690	47.90
SCAB1421	thioesterase	*Streptomyces coelicolor* A3(2), SCO7687	44.80
SCAB1411	aryl-acid activating	*Bacillus subtilis*, DhbE	51.20
SCAB1401	TetR-family regulator	*Streptomyces* sp., unchar. regulator	56.19
SCAB1381	salicylate synthase	*Mycobacterium tuberculosis*, MbtI	41.21
SCAB1371	transcriptional activator	*Streptomyces coelicolor* A3(2), SCO4426/AfsR	40.89

Table 6-3 Coding sequence predictions in the pyochelin-like biosynthesis cluster of *Streptomyces scabies* 87-22 and related proteins in other organisms. Function proposed by similarity to characterised proteins with conserved Pfam domains as supporting evidence.

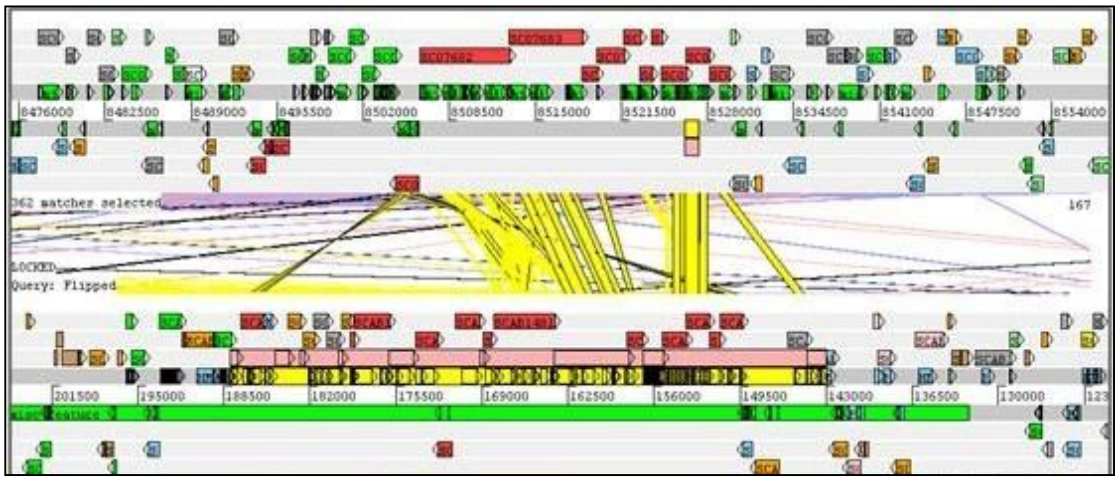


Figure 6-9 Tblastx comparison between "*S. coelicolor*" A3(2) coelibactin cluster (top) and putative pyochelin cluster in *S. scabies* 87.22 (bottom) visualized in Artemis . Alignments bars coloured yellow represent direct and inverted alignments to sequences within the *nrps1* gene cluster SCAB1411-1671.

Similarity with coelibactin cluster

Figure 6-8 (above) shows similarity between coelibactin cluster SCO7681-7683 in "*S. coelicolor*" A3(2) (Bentley *et al.* 2002) and *nrps1* cluster in *S. scabies* 87.22. These clusters appear to be related, and occur as candidate homologues in the orthomcl comparison (In **4.2.2.6).** The cluster in "*S. coelicolor*" A3(2) has an extra biosynthetic module in the NRPS protein, which could indicate a longer product molecule resulting from an extra round of extension on the thiotemplate. The strongest matches amongst the coelibactin cluster are to those coding sequences predicted to encode an export system for the siderophore; the analogous coding sequences in scabies also look likely to encode an export system.

Alternate salicylate supply pathway?

In *Pseudomonas aeruginosa* salicylate is thought to be supplied to the synthase by the action of PchAB (Serino *et al.* 1995). The same researchers demonstrated that PchA catalyses conversion of chorismate to isochorismate and complementation with *pchB* has been shown to allow conversion of isochorismate to salicylate (Serino *et al.* 1995). There is no clear no equivalent to *pchAB* in the S. scabies cluster, so it is predicted that salicylate is acquired otherwise.

SCAB1381 shows similarity to characterised proteins from *Yersinia enterocolitica* (Irp9)(Kerbarh *et al.* 2005; Kerbarh *et al.* 2006) and *Mycobacterium tuberculosis* (MbtI) (Zwahlen *et al.* 2007) which perform salicylate biosynthesis directly from

chorismate, equivalent to the functions of both pchA and pchB. It is therefore suggested that in this cluster salicylate could be supplied by the single protein alternative pathway.

The vital role of salicylate as a plant signaling molecule in pathogen defence (de Torres Zabala *et al.* 2009) raises the possibility of a pathway of acquisition from the plant host ; removal of the SA signal generated by the plant could interrupt SA-mediated defence responses, allowing colonisation to proceed.

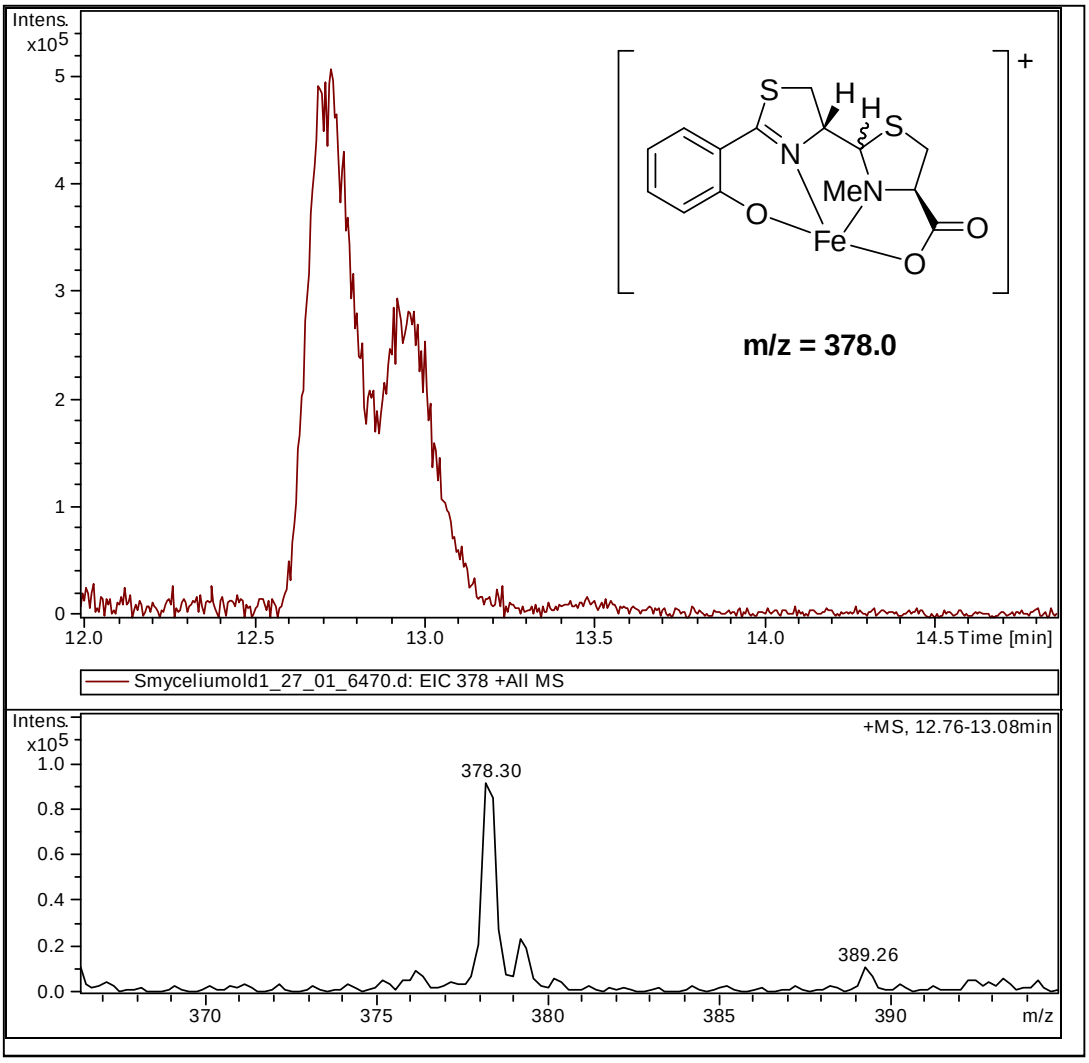


Figure 6-10 Lipid chromatography elution-time graph (above) and time-of-flight mass spectrometry trace (below) are consistent with presence of pyochelin (drawing of primary ion inset), or similar molecule. Supplied by genome project collaborators L. Song and G. L. Challis.

Initial results suggest cluster may be functional

Data from mass spectroscopy suggest pyochelin, or a similar molecule, is produced in small amounts by *S. scabies* 87.22 in iron-deficient liquid medium (G. L. Challis and L. Song, pers. comm.). A peak elutes from liquid chromatography at the expected time for the hydrophobicity of pyochelin (Figure 6-10, top), and this peak has the correct mass for an ion of pyochelin or an isomer (Figure 6-10, bottom trace), as well as having the characteristic three-peaked structure of an iron-binding molecule (G. L. Challis and L. Song, pers. comm.). As iron contains three different mass nuclei in a predictable ratio, so the spectra of molecules with iron bound are easily distinguished by this triple-peaked appearance. Further work could confirm the exact structure of the pyochelin-like molecule. It would be interesting to discover whether production of the pyochelin-like molecule is much different during pathogenicity; or with an exogenous supply of salicylic acid; such as would be available due to the plant's response.

6.2.5 Gene clusters for which there are few clues about the product

6.2.5.1 Cfa-like product? SCAB79601-79721

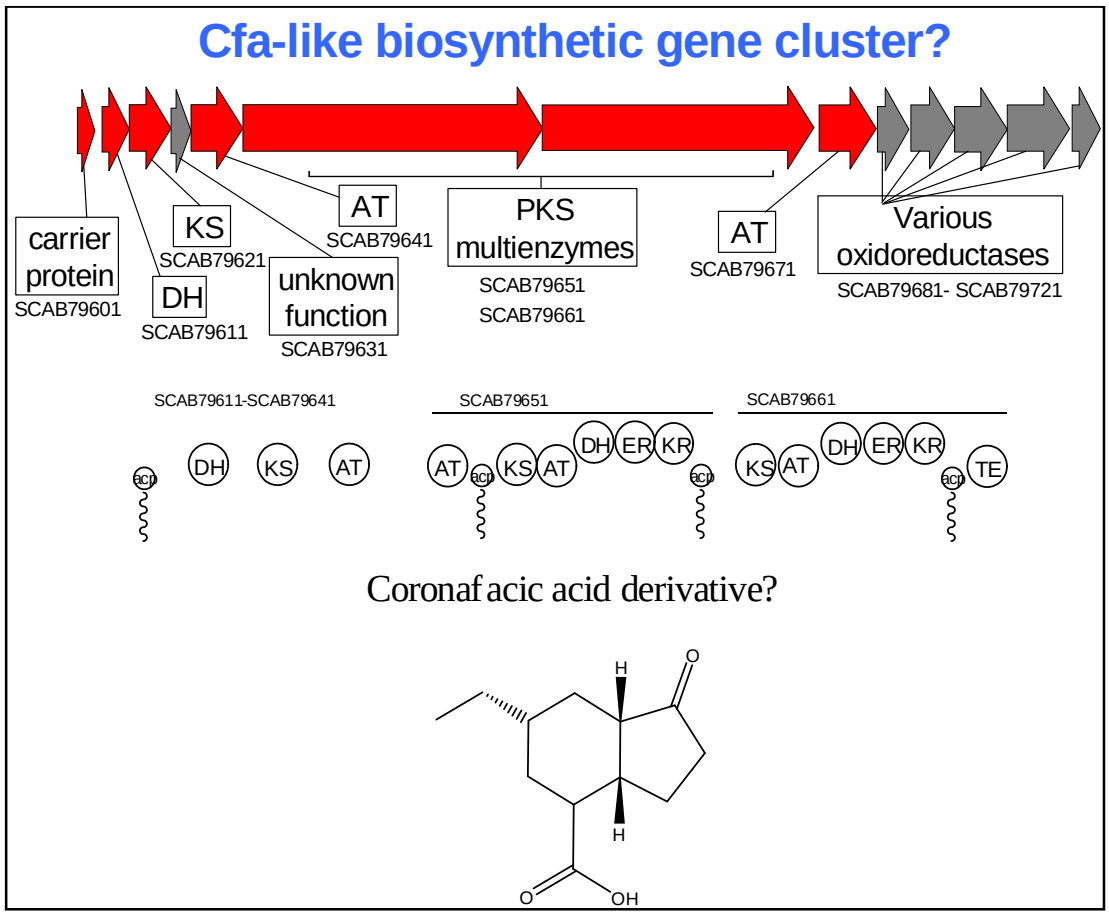


Figure 6-11 Summary of gene cluster possibly producing something related to coronafacic acid.

Thirteen coding sequences were identified (SCAB79601-SCAB79721) including two PKS multienzymes (SCAB79651 and SCAB79661) with similarity to proteins involved in biosynthesis of coronafacic acid, the polyketide component of the hybrid natural product coronatine. Close examination of domains in multienzymes SCAB79651 and SCAB79661 confirm their similarity with equivalents in *Pseudomonas syringae* pv. *glycinea* PG4180 (Liyanage *et al.* 1995), and related cluster sequences with unproven activity in *Pseudomonas syringae* pv. tomato str. DC3000 (Buell *et al.* 2003) and with unproven activity in *Pectobacterium atrosepticum* (previously known as *Erwinia carotovora* subsp. *atroseptica*) SCRI1043 (Bell *et al.* 2004). A gene cluster including polyketide multienzymes was identified in *S. scabies* 87.22 as having some similarity to genes sequenced in *Pseudomonas syringae* pv. *tomato* str. DC3000 and annotated as involved in

biosynthesis of coronafacic acid. A hypothetical biosynthetic scheme based on the functions of the coding sequences discovered in this investigation is appended (Figure 6-14).

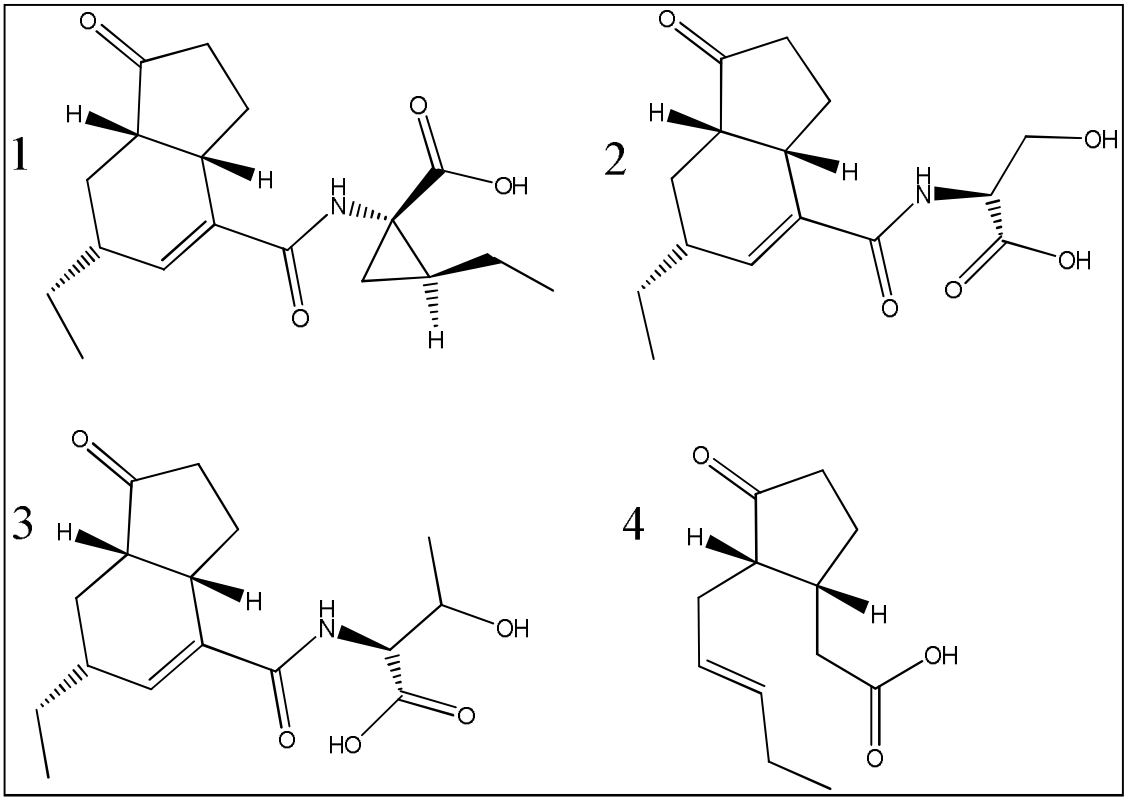


Figure 6-12 Coronatine, related substances produced by phytopathogenic *Pseudomonas syringae* strains, and jasmonic acid, the plant hormone these substances are suspected to resemble. 1, coronatine; **2** N-coronafacyl-L-serine; **3**, N-coronafacyl-L-threonine; **4**, jasmonic acid.

Coronatine is a phytotoxin which contributes substantially to virulence in many pathovars of the microbial plant pathogen *Pseudomonas syringae* (Uppalapati *et al.* 2007). The biosynthetic gene clusters for coronatine biosynthesis are frequently found on a 90 kb self-transmissible plasmid (Alarcon-Chaidez *et al.* 1999). A biosynthesis pathway for the polyketide component coronafacic acid (CFA) has been partly elucidated (Rangaswamy *et al.* 1998; Brooks *et al.* 2004). Biosynthesis appears to be thermoregulated with greatest transcriptional activity at 18°C (Ullrich and Bender 1994), via sensor-regulator CorRSP in *Pseudomonas syringae* pv. *tomato* str. DC3000 (Smirnova *et al.* 2002), but varying with host background (Weingart *et al.* 2004).

Coronatine is composed of a polyketide component, CFA, linked by an amide bond to a peptide acid component, coronamic acid (CMA). This molecule is illustrated (**1**)

in Figure 6-12. It causes chlorosis, hypertrophy, inhibition of root elongation and stimulation of ethene biosynthesis (Bender *et al.* 1996). Sturctural resemblance between the CFA moiety and octadecanoid plant hormones such as jasmonic acid (Figure 6-11 **4**) has been noted by several researchers, and suggests a possible mode of action. Since coronatine has more pathogenic effects than methyl jasmonate or coronafacic acid alone (Palmer and Bender 1995) it is likely that the non-CFA ligated component plays a role; it may be that ligation stabilizes an active conformation.

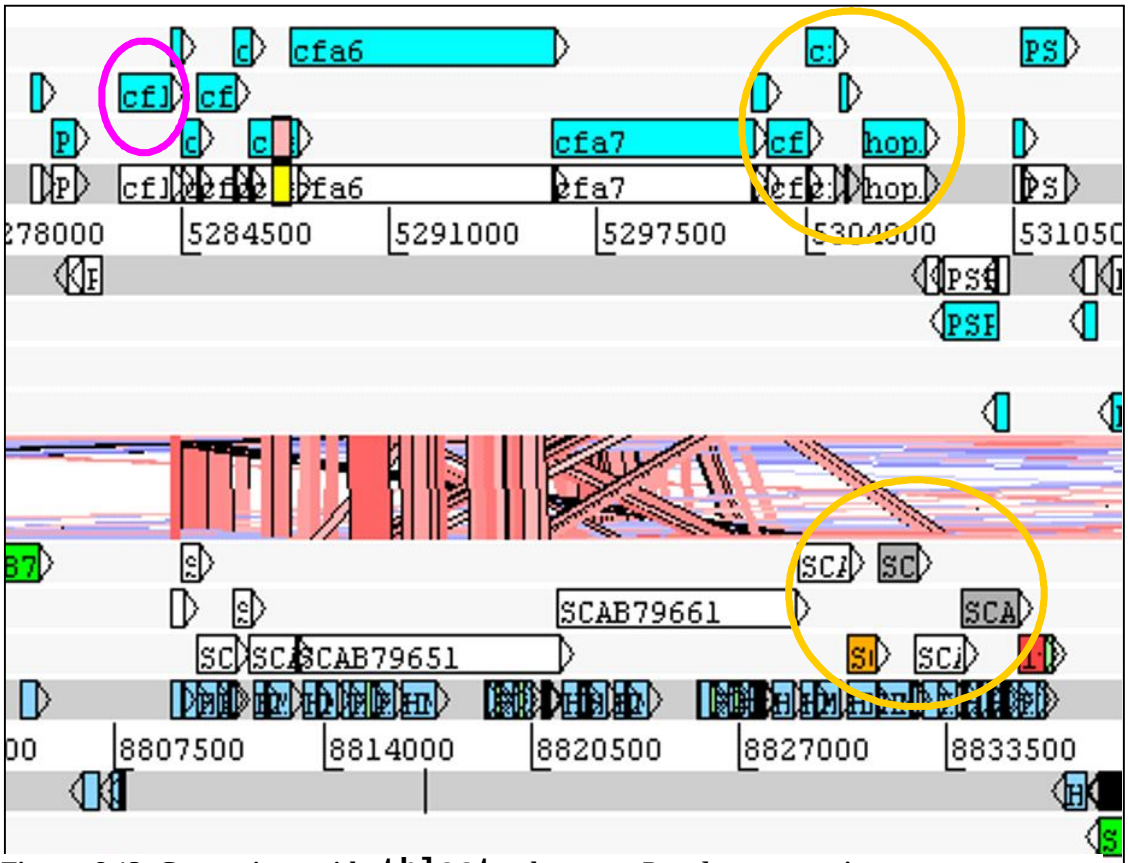


Figure 6-13 Comparison with `tblastx` between *Pseudomonas syringae* pv. tomato str. DC3000 (top) and *Streptomyces scabies* 87-22 (below) visualized in ACT. Notable differences: gold circles highlighting different coding sequences downstream of PKS multienzymes in the two clusters; coding sequence SCAB79661 is longer than cfa7 because of the extra ER domain.

No CMA genes present

The *cma* genes, for biosynthesis of the CMA moiety of coronatine, have not been found in the *S. scabies* 87.22 genome. Other amino acids have been found linked to the CFA component (**2** and **3**, Figure 6-12) in coronatine producers and these compounds also cause chlorosis (Mitchell and Ford 1998). The genes involved in

ligation of CFA to CMA or other amino acids have not yet been identified in any producer. It is possible that *S. scabies* 87.22 biosynthesizes an amino-acid linked coronafacic acid product with phytotoxic or other activity.

Extra ER domain

SCAB79661 is longer than the equivalent multienzyme PKS *cfa7* in the *Pseudomonas syringae* pv *tomato* str. DC3000 cluster, and the extra material appears to be an ER domain. ER domains reduce the double bond in a PKS product left by ketoreduction and dehydration of the incorporated ketide, hence if functional this domain might be expected to reduce the double bond in the product molecule.

Missing thioesterase?

A freestanding thioesterase (Cfa9) is encoded in the gene cluster in *Pseudomonas syringae* pv *tomato* str. DC3000, and may be responsible for releasing the product from PKS multienzyme Cfa6 (Rangaswamy *et al.* 1998). No freestanding thioesterase such as *cfa9* is apparent in the *S. scabies* 87.22 cluster. Freestanding thioesterases have been found to be essential for expression in other complex product biosynthesis clusters (Reimmann *et al.* 2004), perhaps for removal of noncognate molecules from the multienzyme complex, clearing the multienzyme 'assembly line' if it gets blocked by incorrect products.

A coding sequence downstream (SCAB79741) appears to carry the α/β hydrolase fold, related to those in freestanding thioesterases (Holmquist 2000). It is possible that a divergent hydrolase could fulfil the thioesterase function, or it could be encoded elsewhere in the genome. Further investigation could reveal sequences elsewhere in the genome with related expression patterns, for example by temporal patterns of transcription in a microarray study. It is possible that complementation with the *cfa9* coding sequence could restore function, if this cluster were found to be nonfunctional.

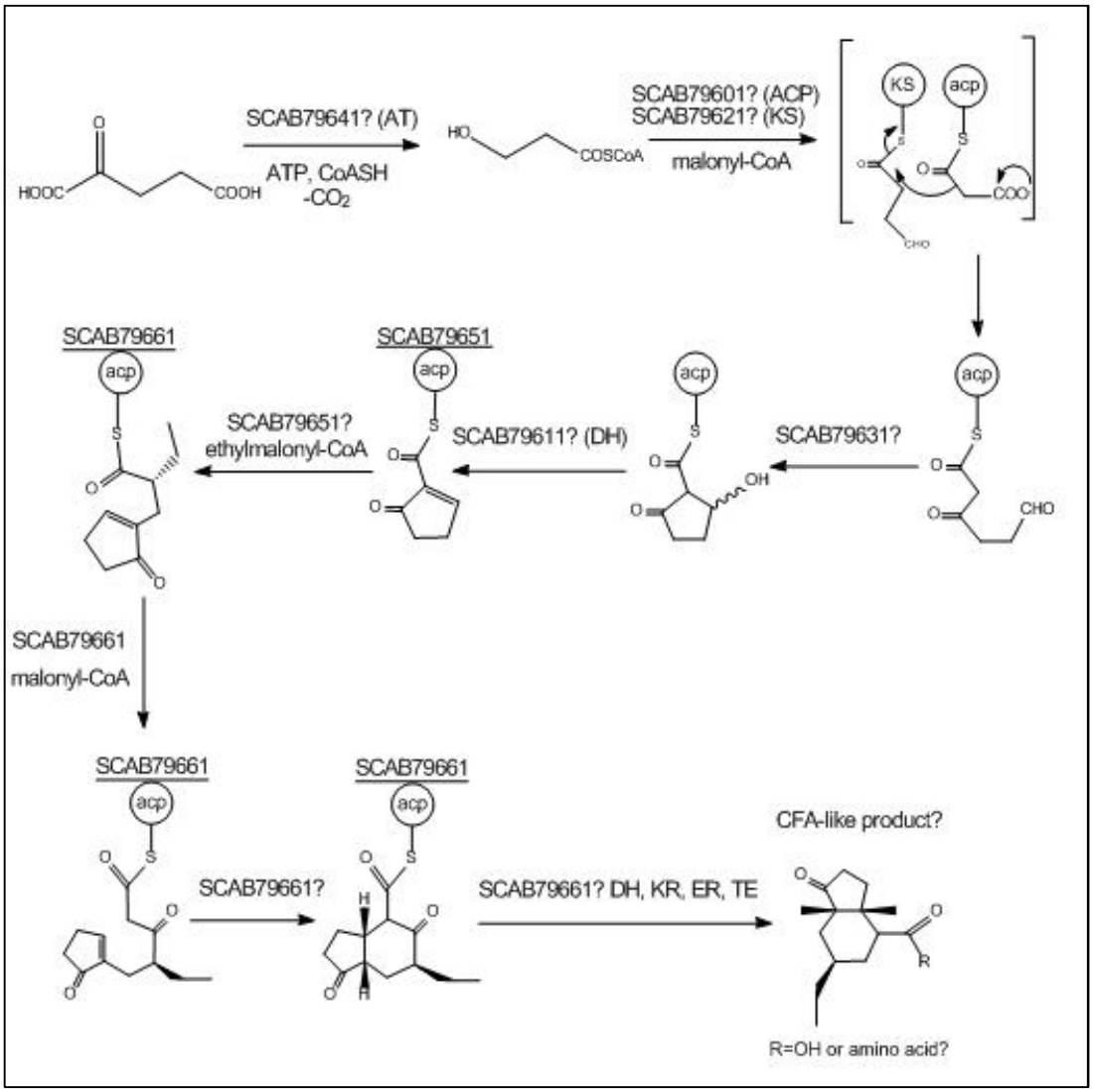


Figure 6-14 Hypothetical biosynthesis scheme for coronafacic-acid-like product in *S. scabies* 87.22 after scheme for coronatine biosynthesis described by Rangaswamy *et al.* 1998.

Similar product, or not?

Besides the differences already mentioned (no *cma* genes, extra ER domain, no *cfa9* thioesterase) between this cluster in *S. scabies* 87.22 and the characterised cluster in *Pseudomonas syringae* pv. *tomato* str. DC3000, there several others. Firstly, the *cfl* coding sequence, necessary for coronatine production in the *Pseudomonas syringae* pv *tomato* str. DC3000 (Wang, X. *et al.* 2002), is not conserved in the *S. scabies* cluster. A coding sequence SCAB79671 with a similar AT domain is found downstream and could possibly fulfil the same function.

Secondly, the *S. scabies* 87.22 cluster may have more capacity to reduce polyketide products. Several of the downstream coding sequences encode enzymes capable of

performing the reductive loop functions: SCAB79701 has a domain typical of DH function, SCAB79711 has domains typical of ketoreduction, and SCAB79721 has a domain often found in ER proteins. Modification of the polyketide assembly process either during type II-like processing through the equivalent proteins to Cfa1234 (see scheme) or *in trans* during type I-like polyketide processing on the multienzymes analogous to Cfa67 (Rangaswamy *et al.* 1998), would alter the structure of the polyketide product.

In addition, there may be modifications after PKS assembly before or after release from the multienzyme. For example, two enzymes encoded the downstream region, SCAB79681 and SCAB79691, belong to a monooxygenase and P450-like protein family which can perform addition of hydroxyl groups in other biosynthetic gene clusters (as does P450-like protein in thaxtomin biosynthesis (Healy *et al.* 2002)).

Just upstream of the *cfa* cluster in *Pseudomonas syringae* pv. *tomato* str. DC3000, coding sequence PSPTO_4679 is annotated as a pseudogene; as these can be identified in association with recombination, it may indicate that the gene cluster has been horizontally acquired by this pseudomonad. The graph of G+C content in this region of the *Pseudomonas syringae* pv. *tomato* str. DC3000 genome is higher than the genome mean value. Streptomycetes are some of the few bacteria with higher G+C content genomes than pseudomonads, perhaps the cluster originated in a streptomycete and subsequently transferred to a gram negative bacterium lineage.

Just upstream of the *cfa* cluster in *Pseudomonas syringae* pv. *tomato* str. DC3000, coding sequence PSPTO_4679 is annotated as a pseudogene; as these can be identified in association with recombination, it may indicate that the gene cluster has been horizontally acquired by this pseudomonad. The graph of G+C content in this region of the *Pseudomonas syringae* pv. *tomato* str. DC3000 genome is higher than the genome mean value. Streptomycetes are some of the few bacteria with higher G+C content genomes than pseudomonads, perhaps the cluster originated in a streptomycete and subsequently transferred to a Gram negative bacterium lineage.

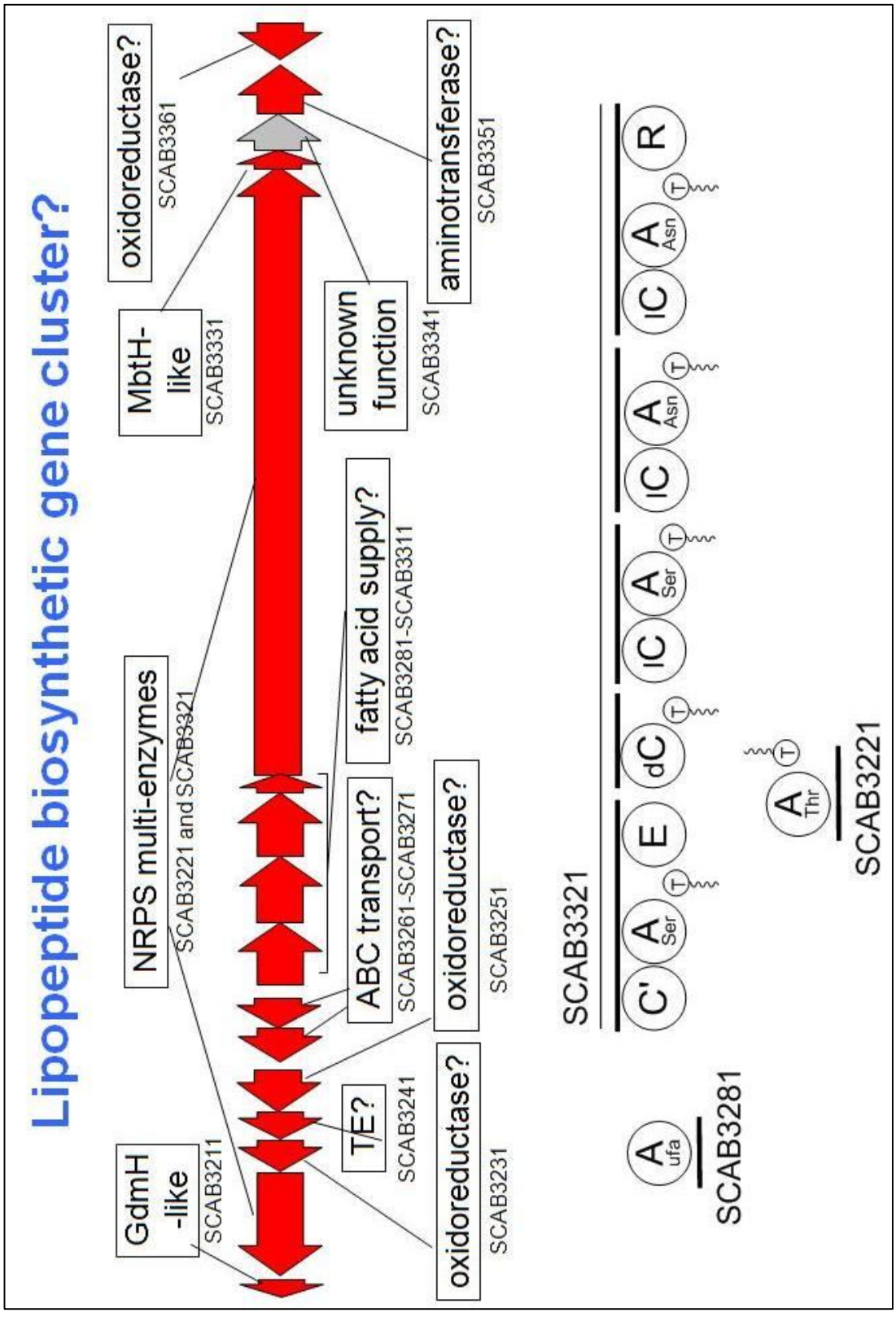


Figure 6-15 Overview of possible lipopeptide gene cluster and proposed modular architecture of nonribosomal peptide synthetase proteins.

6.2.5.2 Lipopeptide SCAB3211-SCAB3361

This gene cluster (Figure 6-12) probably consists of two divergently transcribed operons SCAB3211-SCAB 3261, and SCAB3281-SCAB3351. SCAB3361 could also be involved, since it is a kind of oxidoreductase often found in biosynthetic gene clusters (Table 6-4). It is proposed that an NRPS system and an ABC transport system for secretion of the product molecule are encoded. The four coding sequences SCAB3281-SCAB3311 probably activate, modify and attach an unsaturated fatty acid which is condensed onto the next residue by the 'starter C' condensation domain of SCAB321 multienzyme. The gene for an MbtH homologue is also present in this cluster (SCAB3331) and may function as similar proteins do (see section **6.2.3.1**, **MbtH-like protein TxtH** for more details).

CDS in *S. scabies* 87.22	predicted function
SCAB3211	GdmH-like transport-associated protein
SCAB3221	NRPS multienzyme (AMP-binding and pcp)
SCAB3231	putative oxidoreductase
SCAB3241	putative thioesterase
SCAB3251	putative oxidoreductase
SCAB3261	putative integral membrane transport protein
SCAB3271	putative ABC transporter ATP-binding protein
SCAB3281	putative NRPS-associated AMP-binding protein
SCAB3291	putative oxidoreductase
SCAB3301	putative oxidoreductase
SCAB3311	putative carrier protein for NRPS multienzyme
SCAB3321	putative non-ribosomal peptide synthetase
SCAB3331	MbtH-like protein
SCAB3341	putative oxidoreductase
SCAB3351	putative aminotransferase
SCAB3341	putative NT-sugar reductase

Table 6-4 Summary of predicted functions of coding sequences in the cluster which may encode biosynthetic enzymes producing a lipopeptide.

The other transcription unit of this cluster, beginning with SCAB3271, appears to transport proteins of the ABC family (SCAB3261, SCAB3271). Three coding sequences follow (SCAB3241, SCAB3231, SCAB3221) perhaps involved in modifications and maintenance of the synthetase system. The type II (freestanding) thioesterase SCAB3241 may be involved in removal of non-cognate substrates from the synthetase system. SCAB3211 carries a conserved domain PF02698 (DUF218),

and may encode a Gdm-like protein like those found as accessory factors for ABC transporters involved in secreting antimicrobial peptides (Hille *et al.* 2001), so perhaps comprises part of the ABC transport system.

Specificity of the synthetase domains

The first adenylation domain encoded in coding sequence SCAB3321 is predicted to activate L-serine. Domains activating serine almost invariably show conserved Ser at position 301 (Challis *et al.* 2000). This is not the case for the first adenylation domain in SCAB3321, although the mostly nearly related domains activate serine (Table 6-5). The binding pocket can be predicted to bind a small uncharged polar amino acid using the logic of structure-based prediction methods (Challis et al 2000) and other methods detailed in **3.2.2.1**. This active site has serine at 299 and 322 – could either of those residues might complement the usual function of Ser301, perhaps hydrogen bonding stabilizing the side chain (Challis et al 2000)?

S. scabies	A domain	proposed critical residues of adenylation domain									proposed to	
CDS id	(residues)	235	236	239	278	299	301	322	330	331	activate	nearest
3221	362-762	D	F	W	S	V	G	M	V	H	**L-Thr**	**ABD65960**
3321	486-866	D	V	W	H	S	G	S	V	T	L-Ser	BAE98155
3321	2448-2833	D	V	W	H	V	S	A	V	D	L-Ser	AAY91421
3321	3480-3877	D	L	T	K	I	G	A	V	N	**L-Asn**	**CAB38517**
3321	4518-4918	D	M	T	K	V	G	E	V	G	L-Asn	AAG02354

Table 6-5 Possible critical residues of adenylation domains in cluster thought to encode lipopeptide NRPS system. 'Nearest' indicates INSDC accession of proteins with most similar domains; those listed in bold have identical critical residues; see lookup table Table 2-1 for details of module, protein, and strain. Critical residues of the adenylation domain are referred to by numbering from 1amuA (Conti et al. 1997) as used by previous authors (Stachelhaus et al. 1999; Challis et al. 2000).

The first residue in the peptide chain formed by the SCAB3321 multienzyme is predicted to be incorporated as the D enantiomer, from the condensation and carrier domain specifities. It is proposed that the residue incorporated in the second module is supplied by the protein encoded by SCAB3221. This is probably L-Thr by similarity to EndD in enduracidin biosynthesis, which is known to have this kind of *in trans* supply arrangement.

Although the incorporation in the enduracidin system is L-allo-threonine (**3**: 2S, 3S stereochemistry), other domains with identical critical residues incorporate D-allo (**4**: 2R, 3R) and L-threonine (**1**, the dominant natural enantiomer, 2S, 3R – D-thr **2** is the

mirror image - 2R, 3S) (Yin and Zabriskie 2006). Hence, the different stereoisomers are likely to be produced by a mechanism other than adenylation domain specificity (for example racemisation and selectivity by C-type domains as occurs for normal D enantiomers) so it has been assumed that the incorporation in this *S. scabies* 87.22 gene cluster will be L-Thr.

The next three modules of SCAB3321 probably incorporate L-serine and two L-asparagine residues (Table 6-5) by similarity. The final domain in the multienzyme appears to encode a reductase, hence may releases the product molecule from the synthetase by reduction rather than via a thioesterase. Biosynthesis systems of myxalamid (Silakowski *et al.* 2001), myxochelin (Silakowski *et al.* 2000) and saframycin (Pospiech *et al.* 1996; Li, L. *et al.* 2008) feature terminal reductase domains, and it might be useful to compare the structures of those products if the product of this cluster is sought. The terminal domain of SCAB3321 is most similar to MxaA in myxalamid biosynthesis, and MxcG is the next most nearly related of these three (by blastp vs nr). Several coding sequences for modification enzymes are found in this cluster including a possible aminotransferase SCAB3351, and other predicted products which could add or reduce hydroxyl groups adding further diversity in possible products.

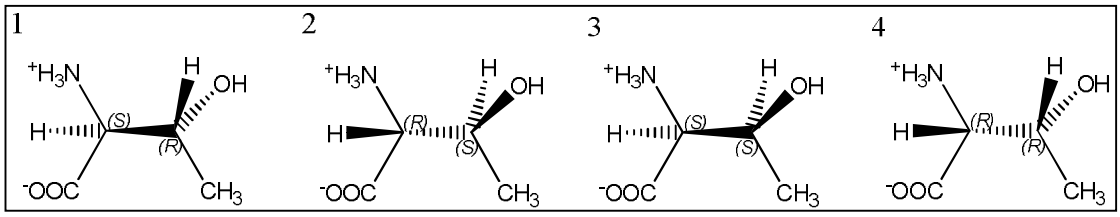


Figure 6-16 Enantiomers of threonine. (Yin et al. 2006.) (1) L-threonine, the dominant natural enantiomer, 2S, 3R; (2) D-thr, the mirror image 2R, 3S; (3) L-allo-threonine with 2S, 3S stereochemistry, other domains with identical critical residues incorporate (4) D-allo threonine 2R, 3R.

6.2.5.3 Peptide siderophore SCAB85431-SCAB85501

This gene cluster appears to consist of ten coding sequences, as judged from the consecutive arrangement of relevant protein products. There may be two transport systems encoded in this gene cluster (SCAB85431-SCAB85451 and SCAB85481-SCAB85501). One of these probably has a secreted binding protein component and may comprise an iron-siderophore uptake transport system. An MbtH homologue is

present in this gene cluster (SCAB85461) and could provide a mechanism for cross talk between pathways for example to co-ordinate production as discussed above (insert cross-reference).

Two coding sequences in this cluster (SCAB85511 formyltransferase and SCAB85521 peptide *N*-oxygenase) are likely to encode enzymes for processing amino acids for incorporation into the product by the multienzyme. This pair of proteins are similar to those found in the coelichelin cluster of "*S. coelicolor*" A3(2) and may function as those do to prepare the N-formyl ornithine residue.

CDS in *S. scabies* 87.22	predicted function	most similar protein in "*S. coelicolor*" A3(2)	amino acids identical (/%)
SCAB85431	ABC transporter integral membrane component	SCO0491	37
SCAB85441	ABC transporter ATP-binding component	SCO0493	46
SCAB85451	secreted siderophore-binding lipoprotein	SCO0494	39
SCAB85461	MbtH homologue	SCO3218	69
SCAB85471	NRPS multienzyme	SCO3230	35
SCAB85481	ABC transporter integral membrane component	SCO1787	51
SCAB85491	ABC transporter integral membrane component	SCO1786	46
SCAB85501	ABC transport system ATP-binding component	SCO1785	55
SCAB85511	formyltransferase	SCO0499	75
SCAB85521	peptide N-oxygenase	SCO0498	60

Table 6-6 Coding sequences in biosynthetic gene cluster predicted to encode a peptide siderophore.

The multienzyme SCAB85471 appears to consist of five modules (Figure 6-12). One or more of the residues activated may be formylated by the product of SCAB85511; N-formyl ornithine is known in some other NRPS systems (Challis and Ravel 2000; Lamont *et al.* 2006). The first adenylation domain has some similarity to a domain activating Phe in thaxtomin biosynthesis of *S. acidiscabies*. Possibly this domain activates N-formyl ornithine which like Phe is a fairly bulky hydrophobic residue.

S. scabies	A domain	proposed critical residues of adenylation domain									proposed to	
CDS id	(residues)	235	236	239	278	299	301	322	330	331	activate	nearest
85471	250-656	D	V	W	I	L	G	A	T	N	?	AAG27088
85471	1721-2146	D	V	W	H	L	S	L	V	D	Ser	**AAC80285**
85471	2804-3208	D	M	E	N	L	G	L	I	N	hOrn	**CAB53322**
85471	4250-4648	D	G	E	D	I	V	L	V	D	Lys/Orn?	AAZ23077
85471	5306-5717	D	A	Q	E	G	G	L	V	D	Orn?	AAX31558

Table 6-7 Proposed critical residues of adenylation domains in this cluster. 'Nearest' indicates INSDC accession of proteins with most similar domains; those listed in bold have identical critical residues. Critical residues of the adenylation domain are referred to by numbering from 1amuA (Conti et al. 1997) as used by previous authors (Stachelhaus et al. 1999; Challis et al. 2000).

It has been suggested that this biosynthetic gene cluster may produce a compound similar or identical to one recently identified in *S. antibioticus* as antichelin (G. L. Challis, pers. comm.). The adenylation domain specificities (Table 2) do not contradict this hypothesis, and the presence of methyltransferase domains in modules 1 and 3 is consistent with it. There is no terminal thioesterase domain in the multienzyme and also no freestanding thioesterase associated with this biosynthetic gene cluster. This is consistent with incorporation of ornithine and may be related to the ease with which ornithine spontaneously cyclises, or forms an amide bond with the C terminus of the peptide chain (as in fengycin, syringomycin, bacitracin (Sieber and Marahiel 2003), which would release a macrocyclic product from the assembly line.

6.2.5.4 Hybrid SCAB78941-SCAB78971

Several coding sequences in this gene cluster appear to be related to those involved in biosynthesis of a lipid virulence factor in *Mycobacteria*: SCAB78961 is related to PpsA and SCAB78971 to FadD26. Diesters of phthiocerol and phenolphthiocerol are involved in protection from macrophage reactive nitrogen intermediates (Rousseau *et al.* 2004) during pathogenicity in *Mycobacterium tuberculosis* (Constant *et al.* 2002). The gene cluster for phthiocerol biosynthesis is considerably larger (Camacho *et al.* 2001) and includes four further PKS multienzymes. Although the similarity is interesting, the distant relationship between these sequences is not directly informative about the structure of a product of the scabies cluster.

SCAB78971 probably activates a saturated fatty acid, by similarity to *M. tuberculosis* FadD26. The presence of a propionyl-coA carboxylase (SCAB789) suggests that the polyketide multienzyme uses methylmalonyl-coA for extension. This PKS/NRPS hybrid cluster is in close proximity to several coding sequences predicted to encode proteins targeted to the cell surface (SCAB78891, SCAB78911, SCAB78921, SCAB78931, SCAB70911). Of these, one (SCAB78931) is related to the cutinase family which are pathogenicity factors in some plant pathogens, and which have also been found in bacterial lineages (Belbahri *et al.* 2008). This family of proteins are involved in breaking down the waxy cuticle on the surface on the aerial parts of plants.

CDS id	function	nearest
SCAB78941	unknown function	no significant matches
SCAB78951	carbohydrate-fatty acid transferase?	AAP23202
SCAB78961	polyketide synthase multienzyme	Q10977 PpsA
SCAB78971	fatty acid adenylation enzyme?	Q10976
SCAB78981	propionyl coA carboxylase?	SCO4926 pccB
SCAB78991	cation binding?	no significant matches

Table 6-8 Coding sequences in hybrid cluster SCAB78941-SCAB78971. Sense indicates orientation on the DNA strand (c = complement or reversed orientation) function reflects product line annotation, nearest describes most closely related characterised protein by blastp vs nr.

6.3 Conclusions

The genome of *S. scabies* 87.22 clearly encodes gene clusters for novel complex products, as well as those already known to be produced by this organism (thaxtomins, concanamycins). It may be that a pyochelin-like siderophore and a product with some similarity to coronafacic acid are produced, and it is possible to envision a role for these substances in pathogenicity. Isolation of natural products is not easy; careful regulation of the clusters means it can be hard to identify the correct conditions for expression. Since a gene cluster for concanamycin product was found in a non-pathogenic streptomycete (Haydock *et al.* 2005), it seems unlikely that concanamycins uniquely contribute to pathogenicity, although there may be some effect.

The presence of three MbtH-like proteins in the thaxtomin, putative lipopeptide, and peptide siderophore clusters could provide a mechanism by which production could be co-ordinated between clusters, as it appears to occur in other organism. Since production of thaxtomins is so directly involved in pathogenicity, perhaps these three clusters are all pathogenicity-associated traits, perhaps expressed when the presence of the plant host is sensed.

The presence of a functioning gene cluster for the production of pyochelin is unexpected, especially since other pyochelin-producing organisms are Gram negative; it raises the question of the origin of such compounds. Further work could study and compare the sequences in depth to form hypotheses about the origins and distribution of such clusters; these findings underline the importance of microbial sequencing projects for comprehension of the origins of virulence factors.

The cluster for a pyochelin-like product deserves further investigation to discover the stereochemistry of the product. It is possible that the biosynthesis pathway encoded in this cluster picks up a supply of salicylic acid (SA) from the plant host's defence response – such a finding could explain why levels produced in iron-free liquid media are very low, and it would be interesting to find out whether supply of additional SA increased production of the pyochelin-like product.

The coronafacic acid-like product predicted, or any of the other clusters described in this chapter, could have roles in pathogenicity. A survey of genomes of scab-causing organisms and non-pathogenic type strains might illuminate how widely these sets of genes are distributed. Given the diverse 16S sequences of scab-causing organisms, conservation of genetic material would be supportive of a role in the pathogenic niche, and further investigations could investigate possible effects on pathogenicity phenotypes.

Any of the gene clusters identified in this study might be entirely silent, non-functional, and hence have no selective advantage. There are suggestions that purifying selection operates efficiently removes genetic material that does not confer selective advantage (Mira *et al.* 2001), so it seems unlikely that all the potential biosynthetic clusters without known activity are silent under all conditions.

6.3.1 Method evaluation

Non-ribosomal peptide synthetase (NRPS) domains

For studies of amino acid adenylation domain, structural models based on the GrsA amino acid adenylation domain (Conti *et al.* 1997) have been used by several groups of researchers investigating the structural basis of specificity (Stachelhaus *et al.* 1999; Challis *et al.* 2000). That model has been used in this work, as used in the structure-based predictive model (Challis *et al.* 2000). More recently a model for the structure of a four-domain module of an NRPS system has been published (Tanovic *et al.* 2008). Detailed comparison between the two models may throw new light on the structure-based modelling process and the limitations of prediction of substrates for this kind of domain.

The model used for alignments of peptidyl carrier proteins (thiolation domains) is an excised PCP domain (PDB: 1DNY), functional for phosphopantetheinylation and adenylation (Stachelhaus *et al.* 1996). It has been assumed in this work that domains in *Streptomyces scabies* 87.22 with close phylogenetic relationship are going to be structurally similar. This could be tested more rigorously using tertiary homology modelling techniques such as those in the Swiss-Model pipeline (Arnold *et al.* 2006). It is assumed that residues shown experimentally to have function in PCP (PDB: 1DNY) and conserved in related domains predicted in *S. scabies* 87.22 have the same function.

Recent bioinformatic investigations in N-methyltransferase conserved domain features (Ansari *et al.* 2008) are likely to be of use in future work studying those domains.

Polyketide domains

Known functional and non-functional dehydratase domains retrieved from INSDC formed separate clades reliably in bootstrapped NJ tree when aligned (not shown), perhaps from a release on constraint which subsequently allows divergence in inactive domains. It is possible this kind of treebuilding could also be used as a fast diagnostic. A crystal structure model of an isolated DH domain from the erythromycin biosynthesis system has recently been published (Keatinge-Clay, A. 2008) (PDB:3EL6) and this may be of use. Although no Pfam A domains seem to match reliably to sequences known to have DH function, several Pfam B domains have strong matches to this crystal structure and may be of use for refining a better search strategy for DH domains.

7 Conclusions

7.1 Capacity for complex natural product biosynthesis

7.1.1 Novel capacity discovered in *S. scabies*

Several clusters encoding enzymes for biosynthesis of complex natural products have been uncovered in this genome. It is likely to be of interest to researchers interested in the undiscovered capacity for producing complex natural products in the *Streptomyces* genus that although only a third of coding sequences predicted are only found in *S. scabies* 87.22 of the three organisms compared, this non-conserved fraction is rich in gene sequences predicted to be involved in biosynthesis of complex natural products. Half of the clusters identified in this work were found to be present only in *S. scabies* 87.22 of the three genomes used for comparison (4-6): biosynthesis of complex natural products seems to be more 'species-specific' than genus-wide, and may be part of the mechanism by which these organisms have diversified to embrace their various niches.

It was suggested at an earlier stage of this work that the gene clusters found to be conserved across the three streptomycete genomes might be essential for life in the soil niche, but production of the membrane stability compound hopene appears not to affect stress resistance phenotype in *S. scabies* 87.22 (Seipke and Loria 2009). Perhaps the physiological role of hopanoids is not yet understood. It may be that the clusters for biosynthetic genes do not have selective advantage, or that the selective advantage operates in a synergistic or contingent fashion as has been suggested for this kind of biosynthesis (Challis and Hopwood 2003).

It has also been suggested that an additional coding sequence is present in the thaxtomins biosynthesis cluster to those previously identified. An MbtH-like protein is predicted, and it is possible this protein would have a role in linking several biosynthetic clusters, possible for co-ordinating their activity (see further **7.3.1.2**).

7.1.2 Limitations of prediction of complex products from sequence data

The in-depth studies of clusters predicted to encode enzymes for complex biosynthesis involve a long chain of inference: which is necessarily only as sound as

the least sound step. If the inferences are robust, predictions are powerful. Inferences can be made more robust by making methods explicit and testable. The model choice for studying each domain in these gene clusters has been guided by the presence of above-threshold matches for models in the Pfam A library (Finn *et al.* 2006; Mulder *et al.* 2007; Coggill *et al.* 2008). An important future direction for this kind of work could be in using a number of additional quantitative methods for structure-based comparison, for example by using techniques for tertiary homology modelling such as the Swiss model pipeline (Arnold *et al.* 2006).

Production or not?

The prediction of encoded capacity for biosynthetic products is at an early stage. The attempt in this work has been to suggest which substances are likely to be possible for the organism to produce. The conditions necessary for actual production of any of these substances may be hard to reproduce in the laboratory, because an extremely large number of possible conditions are already known to trigger production and still many such clusters appear to be silent and their products cryptic. Predictions of encoded capacity for biosynthetic products can only be tentative until demonstrated in laboratory studies.

The products encoded in the *S. scabies* 87.22 genome which have been demonstrated to be produced so far are: thaxtomins (King *et al.* 1989), concanamycins (Natsume *et al.* 2001), melanins (Beausejour and Beaulieu 2004), germicidins and desferrioxamines (as shown by collaborators G. L. Challis and L. Song in association with this work and not yet published) and hopene (Seipke and Loria 2009). Of the others carotenoids, geosmin and spore pigment production are observable by humans so production is assumed.

Amongst the clusters common to *S. scabies*, *"S. coelicolor"* A3(2) and *S. avermitilis* MA-4680, ectoines, and the products of the two shared cryptic NIS systems have not been demonstrated in *S. scabies* 87.22. More of the products only found in *S. scabies* 87.22 of the three remained to be demonstrated (Table 4-6).

Time constraints

The studies presented in this work are constrained by the time available and the large size of the genome. Hypothesis-based investigations of individual gene clusters and their functions are likely to be more fruitful than attempts to provide an overview of the capacity at a genome-wide scale. In-depth investigations of an of these clusters would be most fruitful in close collaboration with laboratory-based biologists. It is useful to have an overview of the entire genome capacity, but in depth studies need further work which will be best with attention to experimental design for testing hypotheses generated through them. It is suggested that the outline investigations in this work are a starting point for studies of the clusters.

7.1.3 Implications of these findings for further discoveries in the genus

Streptomyces genus is known for antibiotic production and most research in the genus and in the wider actinomycete phylogenetic grouping is probably justified on this basis. At this time in history when antibiotics have been overused and resistance is a serious clinical problem, techniques for recovering a greater variety of potential clinical candidates may have vital importance for the future wellbeing of humanity.

S. scabies 87.22 seems to fit into a general trend of approximately 30 clusters with genes encoding complex product biosynthesis. Half of those are not conserved across all three genomes. Since these genomes are in different phylogenetic groups in the *Streptomyces* genus (Williams *et al.* 1983) it is possible that genes for these ten conserved clusters will be found in most of the organism in the genus.

The finding that half of the complex product biosynthetic capacity of an organism such as *S. scabies* 87.22 is unconserved is even more interesting, because it implies that further sequence projects amongst the genus have the potential to reveal many more unknown compounds. Drug discovery is a long and complicated process, but if a great number of candidate bioactive molecules are discovered from environmental organism sequencing, the probability of a few to progress through the entire drug discovery pipeline is increased.

7.1.4 The dereplication problem

Methods for de-replicating gene clusters for complex product biosynthesis have a particular place in the drug discovery process. It is thought that a great many novel substances with potential for therapeutic activity may exist in the wild (Watve *et al.* 2001). The necessity of finding new compounds requires smarter searching. It may be that part of the solution to harvesting the vast diversity of natural products in the environment is by sequencing environmental organisms. If organisms producing substances with demonstrable activity are subjected to sequencing, the entire complement of biosynthetic genes can be acquired to add to the database, since it is likely from the current pattern that several cryptic and silent clusters will be found. Nucleotide probes based on conserved and frequently re-discovered biosynthetic gene clusters could be designed to test whether conserved clusters are found in new organisms to direct sequencing towards organism with proven activity with fewer conserved clusters.

7.1.5 Automation of in-depth studies of clusters

With the ever-increasing volume of sequence data it is important that the new findings from a genome are identified as rapidly as possible. The coding sequences involved in complex product biosynthesis are of great importance because of the huge numbers of therapeutically valuable bioactive compounds identified amongst its members. These sequences can be identified rapidly by using conserved domains models. The studies in this work point the way toward computational methods for automated identification of complex product clusters from new sequence data.

The presence of a certain number of conserved domains within a region of sequence could be used to trigger marking of that sequence region for further study. Conserved domain models using active site residue checking (Mistry *et al.* 2007) could be performed to build module maps of proposed clusters where known domains are found within a certain distance of each other. Conserved architecture of cluster with known biosynthetic clusters from INSDC would serve as a first-pass guess about whether the biosynthetic cluster is conserved or not, for further investigation.

7.2 Annotation - reflections

Visually inspecting and editing thousands of coding sequences provides an opportunity to get a sense of the organism's genetic complement and patterns become apparent which are hard to pick out in other ways. It seems likely that there are betters ways for annotators to become familiar with their work.

7.2.1 Future projects

The increasing volume of sequence data and the variable standards of quality control for that data as submitted to the INSDC obviously have consequences for annotation and similar undertakings which exist in a context shaped by the contents of those databases. The time of computational biologists might be better spent doing the in-depth studies and developing whatever shortcuts and software applications seem possible than producing high quality annotation for complete genome sequences.

7.2.2 Sample pipeline

Genome sequences could be released to INSDC at an early stage, either as draft sequences or as completed sequences with only automated annotation, as has been done with streptomycete sequences recently completed by the Broad Institute - **http://www.broad.mit.edu/annotation/genome/streptomyces_group/News.html**

Even with a great deal of care, it seems impossible to claim that an annotation is free of errors, even more because the front of scientific knowledge is constantly moving. Hence it is hard to argue that the time of genomic analysts should be spent in producing fully curated genomes before public release.

If sequences are finished with all gaps closed an automated annotation pipeline could proceed as follows:

Basic coding sequence prediction

Coding sequence identifiers (stepped to allow additions)

Stable RNA prediction

Automated annotation transfer from trusted sources

Sequence features generated from conserved domain motifs (Interpro)

Sequence features from signal peptide and transmembrane helix detection.

The annotator's time is arguably best spent using a variable-quality automated annotation to answer questions about the content of the genome rather than in improving the annotation of the genome sequence. Detailed annotation can be applied, and coding sequence predictions checked by various methods (**2.5.3**) in regions already identified as likely to be important for the study of the organism

7.2.3 General recommendations

Because of the rapid pace of scientific development it makes sense for comparative studies of genome features to retrieve genomes in a primitive state of annotation, because in order to compare genomes with any confidence identical techniques must be used. So it makes sense for the expert's time to be used in producing interpretations and using the genome information to test hypotheses. Conserved domain searches and similarity-based searches allow genome sequences to be used, in preference to producing high quality annotation over the large expanse of a the complete sequence of a streptomycete.

High quality annotation is clearly necessary as a source for data transfer, but it is unrealistic to expect this from the public databases of sequence data because there are no controls on annotation quality. There is a clear need from the genomic perspective to harvest the knowledge of the research community about the genes in organisms and to transfer this kind of in-depth knowledge into annotation which can be transferred where high levels of similarity make it appropriate.

Hence the time of specialists in particular gene families would be well spent to produce ideal annotation designed for transfer on model gene sequences which have been well investigated. Such specialists can also produce guidance with quantitative indicators of the boundaries for appropriate transfer of that annotation, and can contribute to development of resources such as `Pfam` (Finn *et al.* 2006; Mulder *et al.* 2007; Coggill *et al.* 2008) to produce models for distinguishing functional from non-functional and more greatly diverse sequences. Such data can be shared through

various projects already in existence or published as third party annotation supplied to INSDC.

7.3 *S. scabies* 87.22 the pathogen

Instead of the expected pathogenicity island, known pathogenicity genes are found in at least two locations on the genome (Chapter 5). The *txt* genes for biosynthesis of thaxtomins are diverse when compared with the set in turgidiscabies, implying a long time between divergence of the lineages. The sequences in the other region where known pathogenicity genes occur, around the *nec1* locus, are almost identical to those in the *S. turgidiscabies* Car8 pathogenicity island.

7.3.1 Clues about regulation of pathogenicity

7.3.1.1 Iron box regulation?

Making iron physiologically available is a universal problem for living organisms, and in some pathogens low iron availability is a trigger for activation of pathogenicity traits (Manabe *et al.* 2005). Little was known about iron acquisition in *S. scabies* 87.22 prior to genome sequencing. Complete genome sequencing of the model actinomycete, *"Streptomyces coelicolor" A3(2)*, revealed multiple potential iron scavenging systems (Bentley *et al.* 2002) and has already facilitated several detailed studies of siderophore systems in this genus (Barona-Gomez *et al.* 2004; Lautru *et al.* 2005; Barona-Gomez *et al.* 2006; Lautru *et al.* 2007). Other systems for siderophore production (Fiedler *et al.* 2001), and studies of transport (Bunet *et al.* 2006), regulation (Gunter *et al.* 1993; Gunter-Seeboth and Schupp 1995; Crosa and Walsh 2002; Flores *et al.* 2003; Flores and Martin 2004; Flores *et al.* 2005; Tunca *et al.* 2007) and growth-promoting effects (Yamanaka *et al.* 2005) are starting to illuminate patterns of iron acquisition amongst streptomycetes and across eubacteria.

The target sequences proposed as likely to be activated by low iron availability (**5.2.4**) include several associated with sequences likely to function as regulators of other genes (SCAB19291, SCAB24341). This set of target sequences, likely to be activated by low iron availability include those associated with the siderophore

desferrioxamines (SCAB57921-SCAB57981), a gene cluster identified as likely to encode a non-ribosomal peptide synthetase system (SCAB85461-SCAB85521), and an uncharacterised NIS or IucA/IucC system (SCAB85401-84521). It is possible that further investigation of these sequences could show that they are involved in initiating pathogenicity.

7.3.1.2 MbtH-like protein family?

Three copies of MbtH-like proteins are predicted in this genome. It is possible that these act as this family have been shown to do in other organisms (Lautru *et al.* 2007; Wolpert *et al.* 2007), complementing the activity of the other copies. It has been proposed that these proteins could provide a mechanism for integrating signals between the different biosynthetic pathways that include them (Lautru *et al.* 2007). The three copies in the *S. scabies* 87.22 genome are found in the gene cluster for biosynthesis of thaxtomins (SCAB31771 or *txtH*), the putative lipopeptide (SCAB3331) and the possible peptide siderophore (SCAB85461). Given the importance of thaxtomins in pathogenicity in scabies, a mechanism linking the activity of the two other potential biosynthetic clusters would be interesting. It could be that the products of the other clusters have a role in pathogenicity.

It will be interesting to find out whether other scab disease organisms have the same protein family and the same conserved biosynthetic clusters. SCAB31771 certainly appears to have a homologue in the partial sequences of the *S. turgidiscabies* Car8 pathogenicity island (Kers *et al.* 2005).

7.3.1.3 Other possible mechanisms

There are suggestions that *S. scabies* may have interactions with several plant hormones. Nitric oxide is a plant hormone with roles in defence, also known to be produced by *S. scabies* 87.22 (Johnson *et al.* 2008) as part of thaxtomins biosynthesis. Salicylic acid is another plant hormone which also appears to be generated and utilised by *S. scabies* 87.22: the biosynthetic cluster SCAB1381-SCAB1571 which may produce a substance similar to pyochelin (**6.1.2**) appears not to have the same mechanism for salicylate supply as the gene cluster for pyochelin. This cluster probably has a different mechanism for conversion of chorismate to salicylate. Production of the molecule detected by the collaborators and identified as

possibly pyochelin was at very low levels (G.L.Challis and L. Song); it is interesting to wonder whether further investigation would show more production from this cluster with the supply of exogenous salicylate, as might be acquired after infection of a plant host.

Another possible interaction between *S. scabies* 87.22 and a plant host is provided by the similarity of the biosynthetic cluster identified at SCAB79591-SCAB79721. This cluster may produce a substance with some similarity to coronafacic acid, which is thought to act as a structural mimic of the plant hormone jasmonic acid. In all three of this cases, deletion of the biosynthetic genes in *S. scabies* 87.22 and reinfection of tubers would demonstrate whether the biosynthetic product was essential for pathogenicity; it is more complicated to demonstrate a role in pathogenicity because in typical streptomycete fashion, *S. scabies* probably has several contingent mechanisms to fall back on if one or other is not available.

Besides these the presence of TTA codons which encode a rare tRNA-Leu in "*S. coelicolor*" A3(2) (Takano *et al.* 2003; Hesketh *et al.* 2007) has been noted. The locations of these codons has been written into the annotation file and the locations of a few which are possibly of special interest have been noted .

7.3.2 Future work

Complete sequencing of the genome of *S. scabies* 87.22 is an important step in the journey of understanding scab disease pathogenicity. It should be clear that annotation of the complete genome sequence and this initial study of the data from the sequencing project represent initial steps in comprehension of the genomic basis for pathogenicity traits. It is hoped that this annotation of genome sequence will be a useful resource for future investigators of *S. scabies* 87.22 and for other researchers with an interest in streptomycete genomes.

Bibliography

Abbott, J. C., D. M. Aanensen, K. Rutherford, S. Butcher and B. G. Spratt (2005). WebACT--an online companion for the Artemis Comparison Tool. *Bioinformatics* **21**(18): 3665-6.

Abe, T., T. Ikemura, Y. Ohara, H. Uehara, M. Kinouchi, S. Kanaya, Y. Yamada, A. Muto and H. Inokuchi (2009). tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res* **37**(Database issue): D163-8.

Abremski, K. and S. Gottesman (1982). Purification of the bacteriophage lambda *xis* gene product required for lambda excisive recombination. *J Biol Chem* **257**(16): 9658-62.

Akama, T., K. Suzuki, K. Tanigawa, A. Kawashima, H. Wu, N. Nakata, Y. Osana, Y. Sakakibara and N. Ishii (2009). Whole genome tiling array analysis of *Mycobacterium leprae* RNA reveals high expression of pseudogenes and non-coding regions. *J Bacteriol*.

Alarcon-Chaidez, F. J., A. Penaloza-Vazquez, M. Ullrich and C. L. Bender (1999). Characterization of plasmids encoding the phytotoxin coronatine in *Pseudomonas syringae*. *Plasmid* **42**(3): 210-20.

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman (1990). Basic local alignment search tool. *J Mol Biol* **215**(3): 403-10.

Altschul, S. F. and E. V. Koonin (1998). Iterated profile searches with PSI-BLAST--a tool for discovery in protein databases. *Trends Biochem Sci* **23**(11): 444-7.

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**(17): 3389-402.

Altun, G., W. Zhong, Y. Pan, P. C. Tai and R. W. Harrison (2006). A new seed selection algorithm that maximizes local structural similarity in proteins. *Conf Proc IEEE Eng Med Biol Soc* **1**: 5822-5.

Anderson, A. S. and E. M. Wellington (2001). The taxonomy of *Streptomyces* and related genera. *Int J Syst Evol Microbiol* **51**(Pt 3): 797-814.

Ansari, M. Z., J. Sharma, R. S. Gokhale and D. Mohanty (2008). *In silico* analysis of methyltransferase domains involved in biosynthesis of secondary metabolites. *BMC Bioinformatics* **9**: 454.

Aparicio, J. F., I. Molnar, T. Schwecke, A. Konig, S. F. Haydock, L. E. Khaw, J. Staunton and P. F. Leadlay (1996). Organization of the biosynthetic gene

cluster for rapamycin in *Streptomyces hygroscopicus*: analysis of the enzymatic domains in the modular polyketide synthase. *Gene* **169**(1): 9-16.

Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L. S. Yeh (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**(Database issue): D115-9.

Arnold, K., L. Bordoli, J. Kopp and T. Schwede (2006). The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. *Bioinformatics* **22**(2): 195-201.

Bachi, B. and H. L. Kornberg (1975). Genes involved in the uptake and catabolism of gluconate by *Escherichia coli*. *J Gen Microbiol* **90**(2): 321-35.

Baines, A. L. D., K. Xiao and L. L. Kinkel (2007). Lack of correspondence between genetic and phenotypic groups amongst soil-borne streptomycetes. *Fems Microbiology Ecology* **59**(3): 564-575.

Bairoch, A. (1992). PROSITE: a dictionary of sites and patterns in proteins. *Nucleic Acids Res* **20 Suppl**: 2013-8.

Bao, K. and S. N. Cohen (2003). Recruitment of terminal protein to the ends of *Streptomyces* linear plasmids and chromosomes by a novel telomere-binding protein essential for linear DNA replication. *Genes Dev* **17**(6): 774-85.

Barona-Gomez, F., S. Lautru, F. X. Francou, P. Leblond, J. L. Pernodet and G. L. Challis (2006). Multiple biosynthetic and uptake systems mediate siderophore-dependent iron acquisition in Streptomyces coelicolor A3(2) and Streptomyces ambofaciens ATCC 23877. *Microbiology* **152**(Pt 11): 3355-66.

Barona-Gomez, F., U. Wong, A. E. Giannakopulos, P. J. Derrick and G. L. Challis (2004). Identification of a cluster of genes that directs desferrioxamine biosynthesis in *Streptomyces coelicolor* M145. *J Am Chem Soc* **126**(50): 16282-3.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall and E. L. Sonnhammer (2002). The Pfam protein families database. *Nucleic Acids Res* **30**(1): 276-80.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, R. D. Finn and E. L. Sonnhammer (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res* **27**(1): 260-2.

Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe and E. L. Sonnhammer (2000). The Pfam protein families database. *Nucleic Acids Res* **28**(1): 263-6.

Bateman, A., L. Coin, R. Durbin, R. D. Finn, V. Hollich, S. Griffiths-Jones, A. Khanna, M. Marshall, S. Moxon, E. L. Sonnhammer, D. J. Studholme, C. Yeats and S. R. Eddy (2004). The Pfam protein families database. *Nucleic Acids Res* **32**(Database issue): D138-41.

Bear, I. J. and R. G. Thomas (1964). Nature of Argillaceous Odour. *Nature* **201**: 993-995.

Beausejour, J. and C. Beaulieu (2004). Characterization of *Streptomyces scabies* mutants deficient in melanin biosynthesis. *Canadian Journal of Microbiology* **50**(9): 705-709.

Beausejour, J., C. Goyer, J. Vachon and C. Beaulieu (1999). Production of thaxtomin A by Streptomyces scabies strains in plant extract containing media. *Canadian Journal of Microbiology* **45**(9): 764-768.

Belbahri, L., G. Calmin, F. Mauch and J. O. Andersson (2008). Evolution of the cutinase gene family: evidence for lateral gene transfer of a candidate *Phytophthora* virulence factor. *Gene* **408**(1-2): 1-8.

Bell, K. S., M. Sebaihia, L. Pritchard, M. T. Holden, L. J. Hyman, M. C. Holeva, N. R. Thomson, S. D. Bentley, L. J. Churcher, K. Mungall, R. Atkin, N. Bason, K. Brooks, T. Chillingworth, K. Clark, J. Doggett, A. Fraser, Z. Hance, H. Hauser, K. Jagels, S. Moule, H. Norbertczak, D. Ormond, C. Price, M. A. Quail, M. Sanders, D. Walker, S. Whitehead, G. P. Salmond, P. R. Birch, J. Parkhill and I. K. Toth (2004). Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* subsp. *atroseptica* and characterization of virulence factors. *Proc Natl Acad Sci U S A* **101**(30): 11105-10.

Bender, C. L., D. A. Palmer, A. Penaloza-Vazquez, V. Rangaswamy and M. Ullrich (1996). Biosynthesis of coronatine, a thermoregulated phytotoxin produced by the phytopathogen *Pseudomonas syringae*. *Archives of Microbiology* **166**(2): 71-75.

Bentley, S. D., K. F. Chater, A. M. Cerdeno-Tarraga, G. L. Challis, N. R. Thomson, K. D. James, D. E. Harris, M. A. Quail, H. Kieser, D. Harper, A. Bateman, S. Brown, G. Chandra, C. W. Chen, M. Collins, A. Cronin, A. Fraser, A. Goble, J. Hidalgo, T. Hornsby, S. Howarth, C. H. Huang, T. Kieser, L. Larke, L. Murphy, K. Oliver, S. O'Neil, E. Rabbinowitsch, M. A. Rajandream, K. Rutherford, S. Rutter, K. Seeger, D. Saunders, S. Sharp, R. Squares, S. Squares, K. Taylor, T. Warren, A. Wietzorrek, J. Woodward, B. G. Barrell, J. Parkhill and D. A. Hopwood (2002). Complete genome sequence of the model actinomycete *Streptomyces coelicolor* A3(2). *Nature* **417**(6885): 141-7.

Bentley, S. D. and J. Parkhill (2004). Comparative genomic structure of prokaryotes. *Annu Rev Genet* **38**: 771-792.

Berdy, J. (2005). Bioactive microbial metabolites. *J Antibiot (Tokyo)* **58**(1): 1-26.

Berriman, M. and K. Rutherford (2003). Viewing and annotating sequence data with Artemis. *Brief Bioinform* **4**(2): 124-32.

Bevitt, D. J., J. Cortes, S. F. Haydock and P. F. Leadlay (1992). 6-Deoxyerythronolide-B synthase 2 from *Saccharopolyspora erythraea*.

Cloning of the structural gene, sequence analysis and inferred domain structure of the multifunctional enzyme. *Eur J Biochem* **204**(1): 39-49.

Bibb, M. J., P. R. Findlay and M. W. Johnson (1984). The relationship between base composition and codon usage in bacterial genes and its use for the simple and reliable identification of protein-coding sequences. *Gene* **30**(1-3): 157-66.

Billich, A. and R. Zocher (1987). Enzymatic synthesis of cyclosporin A. *J Biol Chem* **262**(36): 17258-9.

Bisang, C., P. F. Long, J. Cortes, J. Westcott, J. Crosby, A. L. Matharu, R. J. Cox, T. J. Simpson, J. Staunton and P. F. Leadlay (1999). A chain initiation factor common to both modular and aromatic polyketide synthases. *Nature* **401**(6752): 502-5.

Bishop, A., S. Fielding, P. Dyson and P. Herron (2004). Systematic insertional mutagenesis of a streptomycete genome: a link between osmoadaptation and antibiotic production. *Genome Res* **14**(5): 893-900.

Boland, C. A. and W. G. Meijer (2000). The iron dependent regulatory protein IdeR (DtxR) of Rhodococcus equi. *FEMS Microbiol Lett* **191**(1): 1-5.

Bonfield, J. K. and R. Staden (1995). The application of numerical estimates of base calling accuracy to DNA sequencing projects. *Nucleic Acids Res* **23**(8): 1406-10.

Bouarab, K., R. Melton, J. Peart, D. Baulcombe and A. Osbourn (2002). A saponin-detoxifying enzyme mediates suppression of plant defences. *Nature* **418**(6900): 889-892.

Bouchek-Mechiche, K., L. Gardan, P. Normand and B. Jouan (2000). DNA relatedness among strains of *Streptomyces* pathogenic to potato in France: description of three new species, *S. europaeiscabiei* sp. nov. and *S. stelliscabiei* sp. nov. associated with common scab, and *S. reticuliscabiei* sp. nov. associated with netted scab. *Int J Syst Evol Microbiol* **50 Pt 1**: 91-9.

Bramwell, P. A., P. Wiener, A. D. Akkermans and E. M. Wellington (1998). Phenotypic, genotypic and pathogenic variation among streptomycetes implicated in common scab disease. *Lett Appl Microbiol* **27**(5): 255-60.

Braun, P. G., P. D. Hildebrand, T. C. Ells and D. Y. Kobayashi (2001). Evidence and characterization of a gene cluster required for the production of viscosin, a lipopeptide biosurfactant, by a strain of Pseudomonas fluorescens. *Can J Microbiol* **47**(4): 294-301.

Brenner, S. E. (1999). Errors in genome annotation. *Trends Genet* **15**(4): 132-3.

Brooks, D. M., G. Hernandez-Guzman, A. P. Kloek, F. Alarcon-Chaidez, A. Sreedharan, V. Rangaswamy, A. Penaloza-Vazquez, C. L. Bender and B. N. Kunkel (2004). Identification and characterization of a well-defined series

of coronatine biosynthetic mutants of *Pseudomonas syringae* pv. tomato DC3000. *Mol Plant Microbe Interact* **17**(2): 162-74.

Buell, C. R., V. Joardar, M. Lindeberg, J. Selengut, I. T. Paulsen, M. L. Gwinn, R. J. Dodson, R. T. Deboy, A. S. Durkin, J. F. Kolonay, R. Madupu, S. Daugherty, L. Brinkac, M. J. Beanan, D. H. Haft, W. C. Nelson, T. Davidsen, N. Zafar, L. Zhou, J. Liu, Q. Yuan, H. Khouri, N. Fedorova, B. Tran, D. Russell, K. Berry, T. Utterback, S. E. Van Aken, T. V. Feldblyum, M. D'Ascenzo, W. L. Deng, A. R. Ramos, J. R. Alfano, S. Cartinhour, A. K. Chatterjee, T. P. Delaney, S. G. Lazarowitz, G. B. Martin, D. J. Schneider, X. Tang, C. L. Bender, O. White, C. M. Fraser and A. Collmer (2003). The complete genome sequence of the *Arabidopsis* and tomato pathogen *Pseudomonas syringae* pv. tomato DC3000. *Proc Natl Acad Sci U S A* **100**(18): 10181-6.

Bukhalid, R. A., S. Y. Chung and R. Loria (1998). nec1, a gene conferring a necrogenic phenotype, is conserved in plant-pathogenic Streptomyces spp. and linked to a transposase pseudogene. *Mol Plant Microbe Interact* **11**(10): 960-7.

Bukhalid, R. A. and R. Loria (1997). Cloning and expression of a gene from *Streptomyces scabies* encoding a putative pathogenicity factor. *J Bacteriol* **179**(24): 7776-83.

Bukhalid, R. A., T. Takeuchi, D. Labeda and R. Loria (2002). Horizontal transfer of the plant virulence gene, nec1, and flanking sequences among genetically distinct Streptomyces strains in the Diastatochromogenes cluster. *Appl Environ Microbiol* **68**(2): 738-44.

Bunet, R., A. Brock, H. U. Rexer and E. Takano (2006). Identification of genes involved in siderophore transport in *Streptomyces coelicolor* A3(2). *FEMS Microbiol Lett* **262**(1): 57-64.

Bursy, J., A. U. Kuhlmann, M. Pittelkow, H. Hartmann, M. Jebbar, A. J. Pierik and E. Bremer (2008). Synthesis and uptake of the compatible solutes ectoine and 5-hydroxyectoine by *Streptomyces coelicolor* A3(2) in response to salt and heat stresses. *Appl Environ Microbiol* **74**(23): 7286-96.

Buttner, M. J., A. M. Smith and M. J. Bibb (1988). At least three different RNA polymerase holoenzymes direct transcription of the agarase gene (*dagA*) of *Streptomyces coelicolor* A3(2). *Cell* **52**(4): 599-607.

Caffrey, P. (2003). Conserved amino acid residues correlating with ketoreductase stereospecificity in modular polyketide synthases. *Chembiochem* **4**(7): 654-7.

Callister, S. J., L. A. McCue, J. E. Turse, M. E. Monroe, K. J. Auberry, R. D. Smith, J. N. Adkins and M. S. Lipton (2008). Comparative bacterial proteomics: analysis of the core genome concept. *PLoS One* **3**(2): e1542.

Camacho, L. R., P. Constant, C. Raynaud, M. A. Laneelle, J. A. Triccas, B. Gicquel, M. Daffe and C. Guilhot (2001). Analysis of the phthiocerol

dimycocerosate locus of *Mycobacterium tuberculosis*. Evidence that this lipid is involved in the cell wall permeability barrier. *J Biol Chem* **276**(23): 19845-54.

Campbell, A. (2003). Prophage insertion sites. *Res Microbiol* **154**(4): 277-82.

Cane, D. E. and R. M. Watt (2003). Expression and mechanistic analysis of a germacradienol synthase from *Streptomyces coelicolor* implicated in geosmin biosynthesis. *Proc Natl Acad Sci U S A* **100**(4): 1547-51.

Carver, T. J., K. M. Rutherford, M. Berriman, M. A. Rajandream, B. G. Barrell and J. Parkhill (2005). ACT: the Artemis Comparison Tool. *Bioinformatics* **21**(16): 3422-3.

Cerdeno-Tarraga, A. M., A. Efstratiou, L. G. Dover, M. T. Holden, M. Pallen, S. D. Bentley, G. S. Besra, C. Churcher, K. D. James, A. De Zoysa, T. Chillingworth, A. Cronin, L. Dowd, T. Feltwell, N. Hamlin, S. Holroyd, K. Jagels, S. Moule, M. A. Quail, E. Rabbinowitsch, K. M. Rutherford, N. R. Thomson, L. Unwin, S. Whitehead, B. G. Barrell and J. Parkhill (2003). The complete genome sequence and analysis of *Corynebacterium diphtheriae* NCTC13129. *Nucleic Acids Res* **31**(22): 6516-23.

Challis, G. L. (2005). A widely distributed bacterial pathway for siderophore biosynthesis independent of nonribosomal peptide synthetases. *Chembiochem* **6**(4): 601-11.

Challis, G. L. (2008a). Genome Mining for Novel Natural Product Discovery. *Journal of Medicinal Chemistry* **51**(9): 2618-2628.

Challis, G. L. (2008b). Mining microbial genomes for new natural products and biosynthetic pathways. *Microbiology* **154**(Pt 6): 1555-69.

Challis, G. L. and D. A. Hopwood (2003). Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by *Streptomyces* species. *Proc Natl Acad Sci U S A* **100 Suppl 2**: 14555-61.

Challis, G. L. and J. Ravel (2000). Coelichelin, a new peptide siderophore encoded by the *Streptomyces coelicolor* genome: structure prediction from the sequence of its non-ribosomal peptide synthetase. *FEMS Microbiol Lett* **187**(2): 111-4.

Challis, G. L., J. Ravel and C. A. Townsend (2000). Predictive, structure-based model of amino acid recognition by nonribosomal peptide synthetase adenylation domains. *Chem Biol* **7**(3): 211-24.

Chater, K. F. and G. Chandra (2006). The evolution of development in *Streptomyces* analysed by genome comparisons. *FEMS Microbiol Rev* **30**(5): 651-72.

Chen, F., A. J. Mackey, C. J. Stoeckert, Jr. and D. S. Roos (2006). OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* **34**(Database issue): D363-8.

Choi, J. H., H. Y. Jung, H. S. Kim and H. G. Cho (2000). PhyloDraw: a phylogenetic tree drawing system. *Bioinformatics* **16**(11): 1056-8.

Choi, S. S., Y. A. Hur, D. H. Sherman and E. S. Kim (2007). Isolation of the biosynthetic gene cluster for tautomycetin, a linear polyketide T cell-specific immunomodulator from *Streptomyces* sp. CK4412. *Microbiology* **153**(Pt 4): 1095-102.

Choulet, F., B. Aigle, A. Gallois, S. Mangenot, C. Gerbaud, C. Truong, F. X. Francou, C. Fourrier, M. Guerineau, B. Decaris, V. Barbe, J. L. Pernodet and P. Leblond (2006a). Evolution of the terminal regions of the *Streptomyces* linear chromosome. *Mol Biol Evol* **23**(12): 2361-9.

Choulet, F., A. Gallois, B. Aigle, S. Mangenot, C. Gerbaud, C. Truong, F. X. Francou, F. Borges, C. Fourrier, M. Guerineau, B. Decaris, V. Barbe, J. L. Pernodet and P. Leblond (2006b). Intraspecific variability of the terminal inverted repeats of the linear chromosome of *Streptomyces ambofaciens*. *J Bacteriol* **188**(18): 6599-610.

Claus, H. and H. Decker (2006). Bacterial tyrosinases. *Syst Appl Microbiol* **29**(1): 3-14.

Coggill, P., R. D. Finn and A. Bateman (2008). Identifying protein domains with the Pfam database. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 5.

Cole, S. T., K. Eiglmeier, J. Parkhill, K. D. James, N. R. Thomson, P. R. Wheeler, N. Honore, T. Garnier, C. Churcher, D. Harris, K. Mungall, D. Basham, D. Brown, T. Chillingworth, R. Connor, R. M. Davies, K. Devlin, S. Duthoy, T. Feltwell, A. Fraser, N. Hamlin, S. Holroyd, T. Hornsby, K. Jagels, C. Lacroix, J. Maclean, S. Moule, L. Murphy, K. Oliver, M. A. Quail, M. A. Rajandream, K. M. Rutherford, S. Rutter, K. Seeger, S. Simon, M. Simmonds, J. Skelton, R. Squares, S. Squares, K. Stevens, K. Taylor, S. Whitehead, J. R. Woodward and B. G. Barrell (2001). Massive gene decay in the leprosy bacillus. *Nature* **409**(6823): 1007-11.

Combes, P., R. Till, S. Bee and M. C. Smith (2002). The *streptomyces* genome contains multiple pseudo-*attB* sites for the (phi)C31-encoded site-specific recombination system. *J Bacteriol* **184**(20): 5746-52.

Constant, P., E. Perez, W. Malaga, M. A. Laneelle, O. Saurel, M. Daffe and C. Guilhot (2002). Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene. *J Biol Chem* **277**(41): 38148-58.

Conti, E., N. P. Franks and P. Brick (1996). Crystal structure of firefly luciferase throws light on a superfamily of adenylate-forming enzymes. *Structure* **4**(3): 287-98.

Conti, E., T. Stachelhaus, M. A. Marahiel and P. Brick (1997). Structural basis for the activation of phenylalanine in the non-ribosomal biosynthesis of gramicidin S. *Embo J* **16**(14): 4174-83.

Cornish-Bowden, A. (1985). Nomenclature for incompletely specified bases in nucleic acid sequences: recommendations 1984. *Nucleic Acids Res* **13**(9): 3021-30.

Corre, C. and G. L. Challis (2007). Heavy tools for genome mining. *Chem Biol* **14**(1): 7-9.

Crespi, M., D. Vereecke, W. Temmerman, M. Van Montagu and J. Desomer (1994). The *fas* operon of *Rhodococcus fascians* encodes new genes required for efficient fasciation of host plants. *J Bacteriol* **176**(9): 2492-501.

Crosa, J. H. and C. T. Walsh (2002). Genetics and assembly line enzymology of siderophore biosynthesis in bacteria. *Microbiol Mol Biol Rev* **66**(2): 223-49.

Cullen, D. W. and A. K. Lees (2007). Detection of the nec1 virulence gene and its correlation with pathogenicity in Streptomyces species on potato tubers and in soil using conventional and real-time PCR. *J Appl Microbiol* **102**(4): 1082-94.

Dai, W. M., Y. Guan and J. Jin (2005). Structures and total syntheses of the plecomacrolides. *Curr Med Chem* **12**(17): 1947-93.

Daubin, V., E. Lerat and G. Perriere (2003). The source of laterally transferred genes in bacterial genomes. *Genome Biol* **4**(9): R57.

Davelos, A. L., K. Xiao, D. A. Samac, A. P. Martin and L. L. Kinkel (2004). Spatial variation in *Streptomyces* genetic composition and diversity in a prairie soil. *Microbial Ecology* **48**(4): 601-612.

Davis, N. K. and K. F. Chater (1990). Spore colour in *Streptomyces coelicolor* A3(2) involves the developmentally regulated synthesis of a compound biosynthetically related to polyketide antibiotics. *Mol Microbiol* **4**(10): 1679-91.

de Bruijn, I., M. J. de Kock, P. de Waard, T. A. van Beek and J. M. Raaijmakers (2008). Massetolide A biosynthesis in *Pseudomonas fluorescens*. *J Bacteriol* **190**(8): 2777-89.

de Lorenzo, V. and J. B. Neilands (1986). Characterization of *iucA* and *iucC* genes of the aerobactin system of plasmid ColV-K30 in *Escherichia coli*. *J Bacteriol* **167**(1): 350-5.

de Torres Zabala, M., M. H. Bennett, W. H. Truman and M. R. Grant (2009). Antagonism between salicylic and abscisic acid reflects early host-pathogen conflict and moulds plant defence responses. *Plant J*.

Del Vecchio, F., H. Petkovic, S. G. Kendrew, L. Low, B. Wilkinson, R. Lill, J. Cortes, B. A. Rudd, J. Staunton and P. F. Leadlay (2003). Active-site

residue, domain and module swaps in modular polyketide synthases. *J Ind Microbiol Biotechnol* **30**(8): 489-94.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg (1999). Improved microbial gene identification with GLIMMER. *Nucleic Acids Res* **27**(23): 4636-41.

Delepelaire, P. and C. Wandersman (1989). Protease Secretion by *Erwinia chrysanthemi* - Protease-B and Protease-C Are Synthesized and Secreted as Zymogens without a Signal Peptide. *Journal of Biological Chemistry* **264**(15): 9083-9089.

Dellagi, A., M. N. Brisset, J. P. Paulin and D. Expert (1998). Dual role of desferrioxamine in *Erwinia amylovora* pathogenicity. *Mol Plant Microbe Interact* **11**(8): 734-42.

Distler, J., K. Mansouri, G. Mayer, M. Stockmann and W. Piepersberg (1992). Streptomycin biosynthesis and its regulation in Streptomycetes. *Gene* **115**(1-2): 105-11.

Dobrindt, U., F. Agerer, K. Michaelis, A. Janka, C. Buchrieser, M. Samuelson, C. Svanborg, G. Gottschalk, H. Karch and J. Hacker (2003). Analysis of genome plasticity in pathogenic and commensal *Escherichia coli* isolates by use of DNA arrays. *J Bacteriol* **185**(6): 1831-40.

Dobrindt, U., B. Hochhut, U. Hentschel and J. Hacker (2004). Genomic islands in pathogenic and environmental microorganisms. *Nature Reviews Microbiology* **2**(5): 414-424.

Donadio, S., P. Monciardini and M. Sosio (2007). Polyketide synthases and nonribosomal peptide synthetases: the emerging view from bacterial genomics. *Nat Prod Rep* **24**(5): 1073-109.

Donadio, S., M. J. Staver, J. B. McAlpine, S. J. Swanson and L. Katz (1992). Biosynthesis of the erythromycin macrolactone and a rational approach for producing hybrid macrolides. *Gene* **115**(1-2): 97-103.

Doumbou, C. L., V. Akimov, M. Cote, P. M. Charest and C. Beaulieu (2001). Taxonomic study on nonpathogenic streptomycetes isolated from common scab lesions on potato tubers. *Systematic and applied microbiology* **24**(3): 451-456.

Doumbou, C. L., V. V. Akimov and C. Beaulieu (1998). Selection and characterization of microorganisms utilizing thaxtomin A, a phytotoxin produced by *Streptomyces scabies*. *Appl Environ Microbiol* **64**(11): 4313-6.

Drake, E. J., J. Cao, J. Qu, M. B. Shah, R. M. Straubinger and A. M. Gulick (2007). The 1.8 A crystal structure of PA2412, an MbtH-like protein from the pyoverdine cluster of *Pseudomonas aeruginosa*. *J Biol Chem* **282**(28): 20425-34.

Dreier, J. and C. Khosla (2000). Mechanistic analysis of a type II polyketide synthase. Role of conserved residues in the beta-ketoacyl synthase-chain length factor heterodimer. *Biochemistry* **39**(8): 2088-95.

Eckwall, E. C. and J. L. Schottel (1997). Isolation and characterization of an antibiotic produced by the scab disease-suppressive *Streptomyces diastatochromogenes* strain PonSSII. *J Ind Microbiol Biotechnol* **19**(3): 220-5.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**(5): 1792-7.

Edwards, U., T. Rogall, H. Blocker, M. Emde and E. C. Bottger (1989). Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* **17**(19): 7843-53.

Egan, S., P. Wiener, D. Kallifidas and E. M. Wellington (1998). Transfer of streptomycin biosynthesis gene clusters within streptomycetes isolated from soil. *Appl Environ Microbiol* **64**(12): 5061-3.

Egan, S., P. Wiener, D. Kallifidas and E. M. Wellington (2001). Phylogeny of *Streptomyces* species and evidence for horizontal transfer of entire and partial antibiotic gene clusters. *Antonie Van Leeuwenhoek* **79**(2): 127-33.

el-Sayed el, S. A. (2001). Production of thaxtomin a by two species of *Streptomyces* causing potato scab. *Acta Microbiol Immunol Hung* **48**(1): 67-79.

Enright, A. J., S. Van Dongen and C. A. Ouzounis (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**(7): 1575-84.

Essen, S. A., A. Johnsson, D. Bylund, K. Pedersen and U. S. Lundstrom (2007). Siderophore production by *Pseudomonas stutzeri* under aerobic and anaerobic conditions. *Appl Environ Microbiol* **73**(18): 5857-64.

Euzeby, J. P. (1997). List of Bacterial Names with Standing in Nomenclature: a folder available on the Internet. *Int J Syst Bacteriol* **47**(2): 590-2.

Euzeby, J. P. (2009) "List of Prokaryotic names with Standing in Nomenclature."

Expert, D. (1999). Withholding and exchanging iron: Interactions Between *Erwinia* spp. and Their Plant Hosts. *Annu Rev Phytopathol* **37**: 307-334.

Fawaz, F. S. and G. H. Jones (1994). Activation of phenoxazinone synthase expression in *Streptomyces lividans*: characterization of the activator fragment from Streptomyces antibioticus. *Microbiology* **140 (Pt 5)**: 1051-8.

Felsenstein, J. (2008). PHYLIP (Phylogeny Inference Package) version 3.68. Seattle, Distributed by the author.

Fiedler, H. P., P. Krastel, J. Muller, K. Gebhardt and A. Zeeck (2001). Enterobactin: the characteristic catecholate siderophore of Enterobacteriaceae is produced by *Streptomyces* species.(1). *FEMS Microbiol Lett* **196**(2): 147-51.

Finking, R. and M. A. Marahiel (2004). Biosynthesis of nonribosomal peptides. *Annual Review of Microbiology* **58**: 453-488.

Finn, R. D., J. Mistry, B. Schuster-Bockler, S. Griffiths-Jones, V. Hollich, T. Lassmann, S. Moxon, M. Marshall, A. Khanna, R. Durbin, S. R. Eddy, E. L. Sonnhammer and A. Bateman (2006). Pfam: clans, web tools and services. *Nucleic Acids Res* **34**(Database issue): D247-51.

Flardh, K. and M. J. Buttner (2009). *Streptomyces* morphogenetics: dissecting differentiation in a filamentous bacterium. *Nat Rev Microbiol* **7**(1): 36-49.

Flint, S. A., G. Stratigopoulos, A. R. Butler and E. Cundliffe (2002). Expression of *tylM* genes during tylosin production: phantom promoters and enigmatic translational coupling motifs. *J Ind Microbiol Biotechnol* **28**(3): 160-7.

Flores, F. J., C. Barreiro, J. J. Coque and J. F. Martin (2005). Functional analysis of two divalent metal-dependent regulatory genes *dmdR1* and *dmdR2* in *Streptomyces coelicolor* and proteome changes in deletion mutants. *Febs J* **272**(3): 725-35.

Flores, F. J. and J. F. Martin (2004). Iron-regulatory proteins DmdR1 and DmdR2 of *Streptomyces coelicolor* form two different DNA-protein complexes with iron boxes. *Biochem J* **380**(Pt 2): 497-503.

Flores, F. J., J. Rincon and J. F. Martin (2003). Characterization of the iron-regulated *desA* promoter of *Streptomyces pilosus* as a system for controlled gene expression in actinomycetes. *Microb Cell Fact* **2**(1): 5.

Floriano, B. and M. Bibb (1996). *afsR* is a pleiotropic but conditionally required regulatory gene for antibiotic production in *Streptomyces coelicolor* A3(2). *Mol Microbiol* **21**(2): 385-96.

Fry, B. A. and R. Loria (2002). Thaxtomin A: evidence for a plant cell wall target. *Physiological and Molecular Plant Pathology* **60**(1): 1-8.

Galperin, M. Y. and E. V. Koonin (1998). Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol* **1**(1): 55-67.

Gao, B., R. Paramanathan and R. S. Gupta (2006). Signature proteins that are distinctive characteristics of Actinobacteria and their subgroups. *Antonie van Leeuwenhoek* **90**(1): 61-91.

Gardner, P. P., J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy and A. Bateman (2009). Rfam: updates to the RNA families database. *Nucleic Acids Res* **37**(Database issue): D136-40.

Gardy, J. L. and F. S. Brinkman (2006). Methods for predicting bacterial protein subcellular localization. *Nat Rev Microbiol* **4**(10): 741-51.

Gardy, J. L., M. R. Laird, F. Chen, S. Rey, C. J. Walsh, M. Ester and F. S. Brinkman (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**(5): 617-23.

Gascuel, O. and M. Steel (2006). Neighbor-joining revealed. *Mol Biol Evol* **23**(11): 1997-2000.

Gelin, J., J. Mortier, J. Moyroud and A. Chene (1993). Synthetic Studies on Thaxtomin-a and Thaxtomin-B, Phytotoxins Associated with Streptomyces-Scabies, the Causal Organism of Potato Common Scab. *Journal of Organic Chemistry* **58**(13): 3473-3475.

Glauert, A. M. and D. A. Hopwood (1961). The fine structure of *Streptomyces violaceoruber* (*S. coelicolor*). III. The walls of the mycelium and spores. *J Biophys Biochem Cytol* **10**: 505-16.

Gold, B., G. M. Rodriguez, S. A. Marras, M. Pentecost and I. Smith (2001). The *Mycobacterium tuberculosis* IdeR is a dual functional regulator that controls transcription of genes involved in iron acquisition, iron storage and survival in macrophages. *Mol Microbiol* **42**(3): 851-65.

Gordon, D. M. (1992). Rate of plasmid transfer among *Escherichia coli* strains isolated from natural populations. *J Gen Microbiol* **138**(1): 17-21.

Gottesman, M. E. and R. A. Weisberg (1971). In *The Bacteriophage Lambda*. A. D. Hershey. Cold Spring Harbour, N.Y., U.S.A., Cold Spring Harbour Laboratory**:** 113-138.

Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna and S. R. Eddy (2003). Rfam: an RNA family database. *Nucleic Acids Res* **31**(1): 439-41.

Gruss, A. and B. Michel (2001). The replication-recombination connection: insights from genomics. *Current Opinion in Microbiology* **4**(5): 595-601.

Guex, N. and M. C. Peitsch (1997). SWISS-MODEL and the Swiss-PdbViewer: an environment for comparative protein modeling. *Electrophoresis* **18**(15): 2714-23.

Gunter-Seeboth, K. and T. Schupp (1995). Cloning and sequence analysis of the *Corynebacterium diphtheriae* dtxR homologue from *Streptomyces lividans* and *S. pilosus* encoding a putative iron repressor protein. *Gene* **166**(1): 117-9.

Gunter, K., C. Toupet and T. Schupp (1993). Characterization of an iron-regulated promoter involved in desferrioxamine B synthesis in *Streptomyces pilosus*: repressor-binding site and homology to the diphtheria toxin gene promoter. *J Bacteriol* **175**(11): 3295-302.

Guo, J., X. Pu, Y. Lin and H. Leung (2006). Protein subcellular localization based on psi-blast and machine learning. *J Bioinform Comput Biol* **4**(6): 1181-95.

Gust, B., G. L. Challis, K. Fowler, T. Kieser and K. F. Chater (2003). PCR-targeted *Streptomyces* gene replacement identifies a protein domain needed for biosynthesis of the sesquiterpene soil odor geosmin. *Proc Natl Acad Sci U S A* **100**(4): 1541-6.

Haese, A., M. Schubert, M. Herrmann and R. Zocher (1993). Molecular characterization of the enniatin synthetase gene encoding a multifunctional enzyme catalysing N-methyldepsipeptide formation in Fusarium scirpi. *Mol Microbiol* **7**(6): 905-14.

Hara, H., Y. Ohnishi and S. Horinouchi (2009). DNA microarray analysis of global gene regulation by A-factor in *Streptomyces griseus*. *Microbiology*.

Haydock, S. F., J. F. Aparicio, I. Molnar, T. Schwecke, L. E. Khaw, A. Konig, A. F. Marsden, I. S. Galloway, J. Staunton and P. F. Leadlay (1995). Divergent sequence motifs correlated with the substrate specificity of (methyl)malonyl-CoA:acyl carrier protein transacylase domains in modular polyketide synthases. *FEBS Lett* **374**(2): 246-8.

Haydock, S. F., A. N. Appleyard, T. Mironenko, J. Lester, N. Scott and P. F. Leadlay (2005). Organization of the biosynthetic gene cluster for the macrolide concanamycin A in *Streptomyces neyagawaensis* ATCC 27449. *Microbiology* **151**(Pt 10): 3161-9.

Healy, F. G., R. A. Bukhalid and R. Loria (1999). Characterization of an insertion sequence element associated with genetically diverse plant pathogenic *Streptomyces* spp. *Journal of Bacteriology* **181**(5): 1562-1568.

Healy, F. G., S. B. Krasnoff, M. Wach, D. M. Gibson and R. Loria (2002). Involvement of a cytochrome P450 monooxygenase in thaxtomin a biosynthesis by *Streptomyces acidiscabies*. *Journal of Bacteriology* **184**(7): 2019-2029.

Healy, F. G., M. Wach, S. B. Krasnoff, D. M. Gibson and R. Loria (2000). The txtAB genes of the plant pathogen Streptomyces acidiscabies encode a peptide synthetase required for phytotoxin thaxtomin A production and pathogenicity. *Molecular Microbiology* **38**(4): 794-804.

Hesketh, A., G. Bucca, E. Laing, F. Flett, G. Hotchkiss, C. P. Smith and K. F. Chater (2007). New pleiotropic effects of eliminating a rare tRNA from *Streptomyces coelicolor*, revealed by combined proteomic and transcriptomic analysis of liquid cultures. *Bmc Genomics* **8**: -.

Hiard, S., R. Maree, S. Colson, P. A. Hoskisson, F. Titgemeyer, G. P. van Wezel, B. Joris, L. Wehenkel and S. Rigali (2007). PREDetector: a new tool to identify regulatory elements in bacterial genomes. *Biochem Biophys Res Commun* **357**(4): 861-4.

Hill, J. and G. Lazarovits (2005). A mail survey of growers to estimate potato common scab prevalence and economic loss in Canada. *Canadian Journal of Plant Pathology-Revue Canadienne De Phytopathologie* **27**(1): 46-52.

Hille, M., S. Kies, F. Gotz and A. Peschel (2001). Dual role of GdmH in producer immunity and secretion of the Staphylococcal lantibiotics gallidermin and epidermin. *Appl Environ Microbiol* **67**(3): 1380-3.

Hodgson, D. A. (2000). Primary metabolism and its control in streptomycetes: a most unusual group of bacteria. *Adv Microb Physiol* **42**: 47-238.

Holmquist, M. (2000). Alpha/Beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr Protein Pept Sci* **1**(2): 209-35.

Hopwood, D. A. (2006). Soil To Genomics: The *Streptomyces* Chromosome. *Annual Review of Genetics* **40**(1): 1-23.

Horinouchi, S., M. Kito, M. Nishiyama, K. Furuya, S. K. Hong, K. Miyake and T. Beppu (1990). Primary structure of AfsR, a global regulatory protein for secondary metabolite formation in *Streptomyces coelicolor* A3(2). *Gene* **95**(1): 49-56.

Huss, M., G. Ingenhorst, S. Konig, M. Gassel, S. Drose, A. Zeeck, K. Altendorf and H. Wieczorek (2002). Concanamycin A, the specific inhibitor of V-ATPases, binds to the V(o) subunit c. *J Biol Chem* **277**(43): 40544-8.

Hutchings, M. I., P. A. Hoskisson, G. Chandra and M. J. Buttner (2004). Sensing and responding to diverse extracellular signals? Analysis of the sensor kinases and response regulators of *Streptomyces coelicolor* A3(2). *Microbiology* **150**(Pt 9): 2795-806.

Ikeda, H., J. Ishikawa, A. Hanamoto, M. Shinose, H. Kikuchi, T. Shiba, Y. Sakaki, M. Hattori and S. Omura (2003). Complete genome sequence and comparative analysis of the industrial microorganism *Streptomyces avermitilis*. *Nat Biotechnol* **21**(5): 526-31.

Ishikawa, J. and K. Hotta (1999). FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. *FEMS Microbiol Lett* **174**(2): 251-3.

Jiang, J., X. He and D. E. Cane (2006). Geosmin biosynthesis. *Streptomyces coelicolor* germacradienol/germacrene D synthase converts farnesyl diphosphate to geosmin. *J Am Chem Soc* **128**(25): 8128-9.

Johnson, E. G., J. P. Sparks, B. Dzikovski, B. R. Crane, D. M. Gibson and R. Loria (2008). Plant-pathogenic *Streptomyces* species produce nitric oxide synthase-derived nitric oxide in response to host signals. *Chem Biol* **15**(1): 43-50.

Jones, G. H. and D. A. Hopwood (1984). Activation of phenoxazinone synthase expression in *Streptomyces lividans* by cloned DNA sequences from *Streptomyces antibioticus*. *J Biol Chem* **259**(22): 14158-64.

Jones, N. A., S. A. Nepogodiev and R. A. Field (2005). Efficient synthesis of methyl lycotetraoside, the tetrasaccharide constituent of the tomato defence glycoalkaloid alpha-tomatine. *Organic & Biomolecular Chemistry* **3**(17): 3201-3206.

Joshi, M., X. Rong, S. Moll, J. Kers, C. Franco and R. Loria (2007). Streptomyces turgidiscabies secretes a novel virulence protein, Nec1, which facilitates infection. *Mol Plant Microbe Interact* **20**(6): 599-608.

Joshi, M. V., D. R. Bignell, E. G. Johnson, J. P. Sparks, D. M. Gibson and R. Loria (2007). The AraC/XylS regulator TxtR modulates thaxtomin biosynthesis and virulence in *Streptomyces scabies*. *Mol Microbiol* **66**(3): 633-42.

Kadi, N., S. Arbache, L. Song, D. Oves-Costales and G. L. Challis (2008). Identification of a gene cluster that directs putrebactin biosynthesis in *Shewanella* species: PubC catalyzes cyclodimerization of N-hydroxy-N-succinylputrescine. *J Am Chem Soc* **130**(32): 10458-9.

Kadi, N., D. Oves-Costales, F. Barona-Gomez and G. L. Challis (2007). A new family of ATP-dependent oligomerization-macrocyclization biocatalysts. *Nat Chem Biol* **3**(10): 652-6.

Kameoka, D., A. Lezhava, H. Zenitani, K. Hiratsu, M. Kawamoto, K. Goshi, K. Inada, H. Shinkawa and H. Kinashi (1999). Analysis of fusion junctions of circularized chromosomes in *Streptomyces griseus*. *J Bacteriol* **181**(18): 5711-7.

Karlin, S., J. Mrazek and A. M. Campbell (1997). Compositional biases of bacterial genomes and evolutionary implications. *J Bacteriol* **179**(12): 3899-913.

Kataoka, M., K. Ueda, T. Kudo, T. Seki and T. Yoshida (1997). Application of the variable region in 16S rDNA to create an index for rapid species identification in the genus *Streptomyces*. *FEMS Microbiol Lett* **151**(2): 249-55.

Kaup, O., I. Grafen, E. M. Zellermann, R. Eichenlaub and K. H. Gartemann (2005). Identification of a tomatinase in the tomato-pathogenic actinomycete *Clavibacter michiganensis* subsp *michiganensis* NCPPB382. *Molecular Plant-Microbe Interactions* **18**(10): 1090-1098.

Keating, T. A., C. G. Marshall, C. T. Walsh and A. E. Keating (2002). The structure of VibH represents nonribosomal peptide synthetase condensation, cyclization and epimerization domains. *Nat Struct Biol* **9**(7): 522-6.

Keatinge-Clay, A. (2008). Crystal structure of the erythromycin polyketide synthase dehydratase. *J Mol Biol* **384**(4): 941-53.

Keatinge-Clay, A. T., A. A. Shelat, D. F. Savage, S. C. Tsai, L. J. Miercke, J. D. O'Connell, 3rd, C. Khosla and R. M. Stroud (2003). Catalysis, specificity,

and ACP docking site of Streptomyces coelicolor malonyl-CoA:ACP transacylase. *Structure* **11**(2): 147-54.

Keatinge-Clay, A. T. and R. M. Stroud (2006). The structure of a ketoreductase determines the organization of the beta-carbon processing enzymes of modular polyketide synthases. *Structure* **14**(4): 737-48.

Kelemen, G. H., P. Brian, K. Flardh, L. Chamberlin, K. F. Chater and M. J. Buttner (1998). Developmental regulation of transcription of *whiE*, a locus specifying the polyketide spore pigment in *Streptomyces coelicolor* A3 (2). *J Bacteriol* **180**(9): 2515-21.

Kendall, K. J. and S. N. Cohen (1987). Plasmid transfer in *Streptomyces lividans*: identification of a *kil-kor* system associated with the transfer region of pIJ101. *J Bacteriol* **169**(9): 4177-83.

Kerbarh, O., D. Y. Chirgadze, T. L. Blundell and C. Abell (2006). Crystal structures of Yersinia enterocolitica salicylate synthase and its complex with the reaction products salicylate and pyruvate. *J Mol Biol* **357**(2): 524-34.

Kerbarh, O., A. Ciulli, N. I. Howard and C. Abell (2005). Salicylate biosynthesis: overexpression, purification, and characterization of Irp9, a bifunctional salicylate synthase from Yersinia enterocolitica. *J Bacteriol* **187**(15): 5061-6.

Kers, J. A., K. D. Cameron, M. V. Joshi, R. A. Bukhalid, J. E. Morello, M. J. Wach, D. M. Gibson and R. Loria (2005). A large, mobile pathogenicity island confers plant pathogenicity on *Streptomyces* species. *Molecular Microbiology* **55**(4): 1025-1033.

Kers, J. A., M. J. Wach, S. B. Krasnoff, J. Widom, K. D. Cameron, R. A. Bukhalid, D. M. Gibson, B. R. Crane and R. Loria (2004). Nitration of a peptide phytotoxin by bacterial nitric oxide synthase. *Nature* **429**(6987): 79-82.

Kiefer, F., K. Arnold, M. Kunzli, L. Bordoli and T. Schwede (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Res* **37**(Database issue): D387-92.

Kinashi, H., K. Someno and K. Sakaguchi (1984). Isolation and characterization of concanamycins A, B and C. *J Antibiot (Tokyo)* **37**(11): 1333-43.

King, R. R. (1997). Synthesis of thaxtomin C. *Canadian Journal of Chemistry-Revue Canadienne De Chimie* **75**(9): 1172-1173.

King, R. R. and C. H. Lawrence (1995). 4-Nitrotryptophans associated with the *in vitro* production of thaxtomin A by *Streptomyces scabies*. *Phytochemistry* **40**(1): 41-43.

King, R. R., C. H. Lawrence, L. A. Calhoun and J. B. Ristaino (1994). Isolation and Characterization of Thaxtomin-Type Phytotoxins Associated with Streptomyces-Ipomoeae. *Journal of Agricultural and Food Chemistry* **42**(8): 1791-1794.

King, R. R., C. H. Lawrence, M. C. Clark and L. A. Calhoun (1989). Isolation and characterisation of phytotoxins associated with *Streptomyces scabies*. *Journal of the Chemical Society-Chemical Communications*(13): 849-850.

King, R. R., C. H. Lawrence, J. Embleton and L. A. Calhoun (2003). More chemistry of the thaxtomin phytotoxins. *Phytochemistry* **64**(6): 1091-1096.

King, R. R., C. H. Lawrence and J. A. Gray (2001). Herbicidal Properties of the thaxtomin Group of phytotoxins. *Journal of Agricultural and Food Chemistry* **49**(5): 2298-2301.

Komatsu, M., Y. Kuwahara, A. Hiroishi, K. Hosono, T. Beppu and K. Ueda (2003). Cloning of the conserved regulatory operon by its aerial mycelium-inducing activity in an *amfR* mutant of *Streptomyces griseus*. *Gene* **306**: 79-89.

Komatsu, M., H. Takano, T. Hiratsuka, Y. Ishigaki, K. Shimada, T. Beppu and K. Ueda (2006). Proteins encoded by the conservon of *Streptomyces coelicolor* A3(2) comprise a membrane-associated heterocomplex that resembles eukaryotic G protein-coupled regulatory system. *Molecular Microbiology* **62**(6): 1534-1546.

Komatsu, M., M. Tsuda, S. Omura, H. Oikawa and H. Ikeda (2008). Identification and functional analysis of genes controlling biosynthesis of 2-methylisoborneol. *Proc Natl Acad Sci U S A* **105**(21): 7422-7.

Konz, D. and M. A. Marahiel (1999). How do peptide synthetases generate structural diversity? *Chem Biol* **6**(2): R39-48.

Korman, T. P., J. A. Hill, T. N. Vu and S. C. Tsai (2004). Structural analysis of actinorhodin polyketide ketoreductase: cofactor binding and substrate specificity. *Biochemistry* **43**(46): 14529-38.

Koski, L. B., R. A. Morton and G. B. Golding (2001). Codon bias and base composition are poor indicators of horizontally transferred genes. *Mol Biol Evol* **18**(3): 404-12.

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* **305**(3): 567-80.

Kuhnert, P., B. Heyberger-Meyer, A. P. Burnens, J. Nicolet and J. Frey (1997). Detection of RTX toxin genes in Gram-negative bacteria with a set of specific probes. *Appl Environ Microbiol* **63**(6): 2258-65.

Kurtz, S. and C. Schleiermacher (1999). REPuter: fast computation of maximal repeats in complete genomes. *Bioinformatics* **15**(5): 426-7.

Kutzner, H. J. and S. A. Waksman (1959). *Streptomyces coelicolor* Mueller and *Streptomyces violaceoruber* Waksman and Curtis, two distinctly different organisms. *J Bacteriol* **78**: 528-38.

Kyte, J. and R. F. Doolittle (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol* **157**(1): 105-32.

Lally, E. T., R. B. Hill, I. R. Kieba and J. Korostoff (1999). The interaction between RTX toxins and target cells. *Trends Microbiol* **7**(9): 356-61.

Lambalot, R. H., A. M. Gehring, R. S. Flugel, P. Zuber, M. LaCelle, M. A. Marahiel, R. Reid, C. Khosla and C. T. Walsh (1996). A new enzyme superfamily - the phosphopantetheinyl transferases. *Chem Biol* **3**(11): 923-36.

Lambert, D. H. and R. Loria (1989). *Streptomyces scabies* Sp-Nov, Nom-Rev. *International Journal of Systematic Bacteriology* **39**(4): 387-392.

Lambert, D. H., R. Loria, D. P. Labeda and G. S. Saddler (2007). Recommendation for the conservation of the name *Streptomyces scabies*. Request for an opinion. *International Journal of Systematic and Evolutionary Microbiology* **57**: 2447-2448.

Lamont, I. L., L. W. Martin, T. Sims, A. Scott and M. Wallace (2006). Characterization of a gene encoding an acetylase required for pyoverdine synthesis in *Pseudomonas aeruginosa*. *J Bacteriol* **188**(8): 3149-52.

Langille, M. G., W. W. Hsiao and F. S. Brinkman (2008). Evaluation of genomic island predictors using a comparative genomics approach. *BMC Bioinformatics* **9**: 329.

Lapage, S. P., P. H. A. Sneath, E. F. Lessel, V. B. D. Skerman, H. P. R. Seeliger and W. A. Clark (1990). International code of nomenclature of bacteria (1990 revision). Washington, American Society for Microbiology.

Lau, J., H. Fu, D. E. Cane and C. Khosla (1999). Dissecting the role of acyltransferase domains of modular polyketide synthases in the choice and stereochemical fate of extender units. *Biochemistry* **38**(5): 1643-51.

Lautru, S. and G. L. Challis (2004). Substrate recognition by nonribosomal peptide synthetase multi-enzymes. *Microbiology* **150**(Pt 6): 1629-36.

Lautru, S., R. J. Deeth, L. M. Bailey and G. L. Challis (2005). Discovery of a new peptide natural product by Streptomyces coelicolor genome mining. *Nat Chem Biol* **1**(5): 265-9.

Lautru, S., D. Oves-Costales, J. L. Pernodet and G. L. Challis (2007). MbtH-like protein-mediated cross-talk between non-ribosomal peptide antibiotic and siderophore biosynthetic pathways in *Streptomyces coelicolor* M145. *Microbiology* **153**(Pt 5): 1405-12.

Lauzier, A., C. Goyer, L. Ruest, R. Brzezinski, D. L. Crawford and C. Beaulieu (2002). Effect of amino acids on thaxtomin A biosynthesis by Streptomyces scabies. *Canadian Journal of Microbiology* **48**(4): 359-364.

Lawrence, C. H., M. C. Clark and R. R. King (1990). Induction of Common Scab Symptoms in Aseptically Cultured Potato-Tubers by the Vivotoxin, Thaxtomin. *Phytopathology* **80**(7): 606-608.

Lawrence, J. G. and H. Hendrickson (2005). Genome evolution in bacteria: order beneath chaos. *Curr Opin Microbiol* **8**(5): 572-8.

Leblond, P., G. Fischer, F. X. Francou, F. Berger, M. Guerineau and B. Decaris (1996). The unstable region of Streptomyces ambofaciens includes 210 kb terminal inverted repeats flanking the extremities of the linear chromosomal DNA. *Mol Microbiol* **19**(2): 261-71.

Lee, H. S., Y. Ohnishi and S. Horinouchi (2001). A sigmaB-like factor responsible for carotenoid biosynthesis in *Streptomyces griseus*. *J Mol Microbiol Biotechnol* **3**(1): 95-101.

Leiner, R. H., B. A. Fry, D. E. Carling and R. Loria (1996). Probable involvement of thaxtomin A in pathogenicity of Streptomyces scabies on seedlings. *Phytopathology* **86**(7): 709-713.

Li, L., W. Deng, J. Song, W. Ding, Q. F. Zhao, C. Peng, W. W. Song, G. L. Tang and W. Liu (2008). Characterization of the saframycin A gene cluster from *Streptomyces lavendulae* NRRL 11002 revealing a nonribosomal peptide synthetase system for assembling the unusual tetrapeptidyl skeleton in an iterative manner. *J Bacteriol* **190**(1): 251-63.

Li, L., C. J. Stoeckert, Jr. and D. S. Roos (2003). OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**(9): 2178-89.

Li, W. C., J. Wu, W. X. Tao, C. H. Zhao, Y. M. Wang, X. Y. He, G. Chandra, X. F. Zhou, Z. X. Deng, K. F. Chater and M. F. Tao (2007). A genetic and bioinformatic analysis of *Streptomyces coelicolor* genes containing TTA codons, possible targets for regulation by a developmentally significant tRNA. *Fems Microbiology Letters* **266**(1): 20-28.

Lin, Y. S., H. M. Kieser, D. A. Hopwood and C. W. Chen (1993). The chromosomal DNA of *Streptomyces lividans* 66 is linear. *Mol Microbiol* **10**(5): 923-33.

Linne, U., S. Doekel and M. A. Marahiel (2001). Portability of epimerization domain and role of peptidyl carrier protein on epimerization activity in nonribosomal peptide synthetases. *Biochemistry* **40**(51): 15824-34.

Liyanage, H., D. A. Palmer, M. Ullrich and C. L. Bender (1995). Characterisation and transcriptional analysis of the gene cluster for coronafacic acid, the polyketide component of the phytotoxin coronatine. *Applied and Environmental Microbiology* **61**(11): 3843-3848.

Locci, R. (1994). Actinomycetes as plant pathogens. *European Journal of Plant Pathology* **100**(3): 179-200.

Loria, R. (1991) "Vegetable crops - potato scab." *Vegetable MD Online Fact Sheets*,

Loria, R., D. R. Bignell, S. Moll, J. C. Huguet-Tapia, M. V. Joshi, E. G. Johnson, R. F. Seipke and D. M. Gibson (2008). Thaxtomin biosynthesis: the path to plant pathogenicity in the genus *Streptomyces*. *Antonie Van Leeuwenhoek* **94**(1): 3-10.

Loria, R., R. A. Bukhalid, R. A. Creath, R. H. Leiner, M. Olivier and J. C. Steffens (1995). Differential production of thaxtomins by pathogenic *Streptomyces species* in vitro. *Phytopathology* **85**(5): 537-541.

Loria, R., R. A. Bukhalid, B. A. Fry and R. R. King (1997). Plant pathogenicity in the genus *Streptomyces*. *Plant Disease* **81**(8): 836-846.

Loria, R., J. A. Kers and M. V. Joshi (2006). Evolution of plant pathogenicity in *Streptomyces*. *Annual Review of Phytopathology* **44**: 469-487.

Lowe, T. M. and S. R. Eddy (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* **25**(5): 955-64.

Lu, G. and E. N. Moriyama (2004). Vector NTI, a balanced all-in-one sequence analysis suite. *Brief Bioinform* **5**(4): 378-88.

Maddison, W. P. and D. R. Maddison (2009). Mesquite: a modular system for evolutionary analysis Version 2.6.

Mahan, M. J. and J. R. Roth (1988). Reciprocality of recombination events that rearrange the chromosome. *Genetics* **120**(1): 23-35.

Mahan, M. J. and J. R. Roth (1989). Role of recBC function in formation of chromosomal rearrangements: a two-step model for recombination. *Genetics* **121**(3): 433-43.

Manabe, Y. C., C. L. Hatem, A. K. Kesavan, J. Durack and J. R. Murphy (2005). Both *Corynebacterium diphtheriae* DtxR(E177K) and *Mycobacterium tuberculosis* IdeR(D177K) are dominant positive repressors of IdeR-regulated genes in *M. tuberculosis*. *Infection and Immunity* **73**(9): 5988-5994.

Marshall, C. G., N. J. Hillson and C. T. Walsh (2002). Catalytic mapping of the vibriobactin biosynthetic enzyme VibF. *Biochemistry* **41**(1): 244-50.

Martin, J. F. (2004). Phosphate control of the biosynthesis of antibiotics and other secondary metabolites is mediated by the PhoR-PhoP system: an unfinished story. *J Bacteriol* **186**(16): 5197-201.

Martinez-Munoz, G. A. and P. Kane (2008). Vacuolar and plasma membrane proton pumps collaborate to achieve cytosolic pH homeostasis in yeast. *J Biol Chem* **283**(29): 20309-19.

Matoba, Y., T. Kumagai, A. Yamamoto, H. Yoshitsu and M. Sugiyama (2006). Crystallographic evidence that the dinuclear copper center of tyrosinase is flexible during catalysis. *J Biol Chem* **281**(13): 8981-90.

May, J. J., N. Kessler, M. A. Marahiel and M. T. Stubbs (2002). Crystal structure of DhbE, an archetype for aryl acid activating domains of modular nonribosomal peptide synthetases. *Proc Natl Acad Sci U S A* **99**(19): 12120-5.

McClerren, A. L., L. E. Cooper, C. Quan, P. M. Thomas, N. L. Kelleher and W. A. van der Donk (2006). Discovery and *in vitro* biosynthesis of haloduracin, a two-component lantibiotic. *Proc Natl Acad Sci U S A* **103**(46): 17243-8.

McLean, M. J., K. H. Wolfe and K. M. Devine (1998). Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution* **47**(6): 691-696.

Mira, A., H. Ochman and N. A. Moran (2001). Deletional bias and the evolution of bacterial genomes. *Trends Genet* **17**(10): 589-96.

Mistry, J., A. Bateman and R. D. Finn (2007). Predicting active site residue annotations in the Pfam database. *BMC Bioinformatics* **8**: 298.

Mitchell, R. E. and K. L. Ford (1998). Chlorosis-inducing products from *Pseudomonas syringae* pathovars: new N-coronafacoyl compounds. *Phytochemistry* **49**(6): 1579-1583.

Mochizuki, S., K. Hiratsu, M. Suwa, T. Ishii, F. Sugino, K. Yamada and H. Kinashi (2003). The large linear plasmid pSLA2-L of Streptomyces rochei has an unusually condensed gene organization for secondary metabolism. *Mol Microbiol* **48**(6): 1501-10.

Morningstar, A., W. H. Gaze, S. Tolba and E. M. H. Wellington (2006). Evolving gene clusters in soil bacteria. In *Prokaryotic Diversity, Mechanism and Significance*. N. A. Logan, H. M. Lappin-Scott and P. C. F. Oyston. Cambridge, Cambridge University Press**:** 201-222

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, V. Buillard, L. Cerutti, R. Copley, E. Courcelle, U. Das, L. Daugherty, M. Dibley, R. Finn, W. Fleischmann, J. Gough, D. Haft, N. Hulo, S. Hunter, D. Kahn, A. Kanapin, A. Kejariwal, A. Labarga, P. S. Langendijk-Genevaux, D. Lonsdale, R. Lopez, I. Letunic, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, A. N. Nikolskaya, S. Orchard, C. Orengo, R. Petryszak, J. D. Selengut, C. J. Sigrist, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats (2007). New developments in the InterPro database. *Nucleic Acids Res* **35**(Database issue): D224-8.

Mun, H. S., E. J. Oh, H. J. Kim, K. H. Lee, Y. H. Koh, C. J. Kim, J. W. Hyun and B. J. Kim (2007). Differentiation of *Streptomyces* spp. which cause potato scab disease on the basis of partial *rpoB* gene sequences. *Syst Appl Microbiol* **30**(5): 401-7.

Mural, R. J. (2000). ARTEMIS: a tool for displaying and annotating DNA sequence. *Brief Bioinform* **1**(2): 199-200.

Nakamura, Y., T. Gojobori and T. Ikemura (1998). Codon usage tabulated from the international DNA sequence databases. *Nucleic Acids Res* **26**(1): 334.

Nakamura, Y., T. Gojobori and T. Ikemura (2000). Codon usage tabulated from international DNA sequence databases: status for the year 2000. *Nucleic Acids Res* **28**(1): 292.

Natsume, M., M. Taki, N. Tashiro and H. Abe (2001). Phytotoxin Production and Aerial Mycelium Formation by *Streptomyces scabies* and *S. acidiscabies in vitro*. *Journal of General Plant Pathology* **67**(4): 299-302.

Needleman, S. B. and C. D. Wunsch (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol* **48**(3): 443-53.

Nielsen, H., S. Brunak and G. von Heijne (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Eng* **12**(1): 3-9.

Nolan, E. M. and C. T. Walsh (2009). How Nature Morphs Peptide Scaffolds into Antibiotics. *Chembiochem* **10**(1): 34-53.

Norman, C. S. (1964). The Treatment of Iron Overload with Desferrioxamine B. *Ir J Med Sci* **38**: 13-8.

Novakova, R., J. Bistakova and J. Kormanec (2004). Characterization of the polyketide spore pigment cluster *whiESa* in *Streptomyces aureofaciens* CCM3239. *Arch Microbiol* **182**(5): 388-95.

Ogata, H., S. Goto, K. Sato, W. Fujibuchi, H. Bono and M. Kanehisa (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* **27**(1): 29-34.

Ohnishi, Y., J. Ishikawa, H. Hara, H. Suzuki, M. Ikenoya, H. Ikeda, A. Yamashita, M. Hattori and S. Horinouchi (2008). Genome sequence of the streptomycin-producing microorganism *Streptomyces griseus* IFO 13350. *J Bacteriol* **190**(11): 4050-60.

Omura, S., H. Ikeda, J. Ishikawa, A. Hanamoto, C. Takahashi, M. Shinose, Y. Takahashi, H. Horikawa, H. Nakazawa, T. Osonoe, H. Kikuchi, T. Shiba, Y. Sakaki and M. Hattori (2001). Genome sequence of an industrial microorganism *Streptomyces avermitilis*: deducing the ability of producing secondary metabolites. *Proc Natl Acad Sci U S A* **98**(21): 12215-20.

Osborn, A. M. and D. Boltner (2002). When phage, plasmids, and transposons collide: genomic islands, and conjugative- and mobilizable-transposons as a mosaic continuum. *Plasmid* **48**(3): 202-212.

Osbourn, A. (1996). Saponins and plant defence - A soap story. *Trends in Plant Science* **1**(1): 4-9.

Palmer, D. A. and C. L. Bender (1995). Ultrastructure of tomato leaf tissue treated with the Pseudomonad phytotoxin coronatine and comparison with methyl jasmonate. *Molecular Plant-Microbe Interactions* **8**(5): 683-692.

Pan, Y., G. Liu, H. Yang, Y. Tian and H. Tan (2009). The pleiotropic regulator AdpA-L directly controls the pathway-specific activator of nikkomycin biosynthesis in *Streptomyces ansochromogenes. Mol Microbiol*.

Park, D. H., Y. M. Yu, J. S. Kim, J. M. Cho, J. H. Hur and C. K. Lim (2003). Characterization of streptomycetes causing potato common scab in Korea. *Plant Disease* **87**(11): 1290-1296.

Parkhill, J. (2002). Annotation of microbial genomes. In *Methods in Microbiology* B. Wren and N. Dorrell. London, Academic Press**:** 3-26.

Patel, H. M. and C. T. Walsh (2001). *In vitro* reconstitution of the *Pseudomonas aeruginosa* nonribosomal peptide synthesis of pyochelin: characterization of backbone tailoring thiazoline reductase and N-methyltransferase activities. *Biochemistry* **40**(30): 9023-31.

Pearson, W. R. and D. J. Lipman (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A* **85**(8): 2444-8.

Petersen, F., H. Zahner, J. W. Metzger, S. Freund and R. P. Hummel (1993). Germicidin, an autoregulative germination inhibitor of Streptomyces viridochromogenes NRRL B-1551. *J Antibiot (Tokyo)* **46**(7): 1126-38.

Philippe, H. and C. J. Douady (2003). Horizontal gene transfer and phylogenetics. *Current Opinion in Microbiology* **6**(5): 498-505.

Pieper, R., A. Haese, W. Schroder and R. Zocher (1995). Arrangement of catalytic sites in the multifunctional enzyme enniatin synthetase. *Eur J Biochem* **230**(1): 119-26.

Piepersberg, W. (2002). Engogenous Antimicrobial Molecules: An Ecological Perspective. In *Molecular Medical Microbiology*. M. Sussman. San Diego, San Francisco, New York, Boston, London, Sydney, Tokyo, Hartcourt - Academic Press**:** 561-584.

Poralla, K., G. Muth and T. Hartner (2000). Hopanoids are formed during transition from substrate to aerial hyphae in *Streptomyces coelicolor* A3(2). *FEMS Microbiol Lett* **189**(1): 93-5.

Pospiech, A., J. Bietenhader and T. Schupp (1996). Two multifunctional peptide synthetases and an O-methyltransferase are involved in the biosynthesis of the DNA-binding antibiotic and antitumour agent saframycin Mx1 from *Myxococcus xanthus. Microbiology* **142 (Pt 4)**: 741-6.

Possoz, C., C. Ribard, J. Gagnat, J. L. Pernodet and M. Guerineau (2001). The integrative element pSAM2 from *Streptomyces*: kinetics and mode of conjugal transfer. *Mol Microbiol* **42**(1): 159-66.

Pranata, J. and W. L. Jorgensen (1991). Monte Carlo simulations yield absolute free energies of binding for guanine-cytosine and adenine-uracil base pairs in chloroform. *Tetrahedron* **47**(14-15): 2491-2501.

Price-Whelan, A., L. E. P. Dietrich and D. K. Newman (2006). Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nat Chem Biol* **2**(2): 71-78.

Pride, D. T., R. J. Meinersmann, T. M. Wassenaar and M. J. Blaser (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**(2): 145-58.

Pridham, T. G. and A. J. Lyons (1965). Further Taxonomic Studies on Straight to Flexuous Streptomycetes. *J Bacteriol* **89**: 331-42.

Quadri, L. E., J. Sello, T. A. Keating, P. H. Weinreb and C. T. Walsh (1998). Identification of a *Mycobacterium tuberculosis* gene cluster encoding the biosynthetic enzymes for assembly of the virulence-conferring siderophore mycobactin. *Chem Biol* **5**(11): 631-45.

Quevillon, E., V. Silventoinen, S. Pillai, N. Harte, N. Mulder, R. Apweiler and R. Lopez (2005). InterProScan: protein domains identifier. *Nucleic Acids Res* **33**(Web Server issue): W116-20.

Rangaswamy, V., S. Jiralerspong, R. Parry and C. L. Bender (1998). Biosynthesis of the *Pseudomonas* polyketide coronafacic acid requires monofunctional and multifunctional polyketide synthase proteins. *Proc Natl Acad Sci U S A* **95**(26): 15469-74.

Rausch, C., I. Hoof, T. Weber, W. Wohlleben and D. H. Huson (2007). Phylogenetic analysis of condensation domains in NRPS sheds light on their functional evolution. *BMC Evol Biol* **7**: 78.

Raymer, G., J. M. Willard and J. L. Schottel (1990). Cloning, sequencing, and regulation of expression of an extracellular esterase gene from the plant pathogen Streptomyces scabies. *J Bacteriol* **172**(12): 7020-6.

Redenbach, M., F. Flett, W. Piendl, I. Glocker, U. Rauland, O. Wafzig, R. Kliem, P. Leblond and J. Cullum (1993). The *Streptomyces lividans* 66 chromosome contains a 1 MB deletogenic region flanked by two amplifiable regions. *Mol Gen Genet* **241**(3-4): 255-62.

Reeves, C. D., S. Murli, G. W. Ashley, M. Piagentini, C. R. Hutchinson and R. McDaniel (2001). Alteration of the substrate specificity of a modular polyketide synthase acyltransferase domain through site-specific mutations. *Biochemistry* **40**(51): 15464-70.

Reimmann, C., H. M. Patel, L. Serino, M. Barone, C. T. Walsh and D. Haas (2001). Essential PchG-dependent reduction in pyochelin biosynthesis of *Pseudomonas aeruginosa*. *J Bacteriol* **183**(3): 813-20.

Reimmann, C., H. M. Patel, C. T. Walsh and D. Haas (2004). PchC thioesterase optimizes nonribosomal biosynthesis of the peptide siderophore pyochelin in Pseudomonas aeruginosa. *J Bacteriol* **186**(19): 6367-73.

Reimmann, C., L. Serino, M. Beyeler and D. Haas (1998). Dihydroaeruginoic acid synthetase and pyochelin synthetase, products of the *pchEF* genes, are induced by extracellular pyochelin in *Pseudomonas aeruginosa*. *Microbiology* **144 (Pt 11)**: 3135-48.

Rigali, S., H. Nothaft, E. E. Noens, M. Schlicht, S. Colson, M. Muller, B. Joris, H. K. Koerten, D. A. Hopwood, F. Titgemeyer and G. P. van Wezel (2006). The sugar phosphotransferase system of Streptomyces coelicolor is regulated by the GntR-family regulator DasR and links N-acetylglucosamine metabolism to the control of development. *Mol Microbiol* **61**(5): 1237-51.

Rigali, S., M. Schlicht, P. Hoskisson, H. Nothaft, M. Merzbacher, B. Joris and F. Titgemeyer (2004). Extending the classification of bacterial transcription factors beyond the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships. *Nucleic Acids Res* **32**(11): 3418-26.

Rogers, M. F. and A. Ben-Hur (2009). The use of gene ontology evidence codes in preventing classifier assessment bias. *Bioinformatics* **25**(9): 1173-7.

Roongsawang, N., K. Hase, M. Haruki, T. Imanaka, M. Morikawa and S. Kanaya (2003). Cloning and characterization of the gene cluster encoding arthrofactin synthetase from *Pseudomonas* sp. MIS38. *Chem Biol* **10**(9): 869-80.

Rousseau, C., N. Winter, E. Pivert, Y. Bordat, O. Neyrolles, P. Ave, M. Huerre, B. Gicquel and M. Jackson (2004). Production of phthiocerol dimycocerosates protects *Mycobacterium tuberculosis* from the cidal activity of reactive nitrogen intermediates produced by macrophages and modulates the early immune response to infection. *Cell Microbiol* **6**(3): 277-87.

Rutherford, K., J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream and B. Barrell (2000). Artemis: sequence visualization and annotation. *Bioinformatics* **16**(10): 944-5.

Saitou, N. and M. Nei (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**(4): 406-25.

Salyers, A. A., N. B. Shoemaker, A. M. Stevens and L. Y. Li (1995). Conjugative Transposons - an Unusual and Diverse Set of Integrated Gene-Transfer Elements. *Microbiological Reviews* **59**(4): 579-&.

Salzberg, S. L. and A. L. Delcher (2004). Tools for gene finding and whole genome comparison. In *Microbial genomes*. C. M. Fraser, T. D. Read and K. E. Nelson. Totowa, NJ, Humana Press Inc**:** 19-32.

Salzberg, S. L., A. L. Delcher, S. Kasif and O. White (1998). Microbial gene identification using interpolated Markov models. *Nucleic Acids Res* **26**(2): 544-8.

Samel, S. A., G. Schoenafinger, T. A. Knappe, M. A. Marahiel and L. O. Essen (2007). Structural and functional insights into a peptide bond-forming bidomain from a nonribosomal peptide synthetase. *Structure* **15**(7): 781-92.

Samel, S. A., B. Wagner, M. A. Marahiel and L. O. Essen (2006). The thioesterase domain of the fengycin biosynthesis cluster: a structural base for the macrocyclization of a non-ribosomal lipopeptide. *J Mol Biol* **359**(4): 876-89.

Schaerlaekens, K., M. Schierova, E. Lammertyn, N. Geukens, J. Anne and L. Van Mellaert (2001). Twin-arginine translocation pathway in Streptomyces lividans. *J Bacteriol* **183**(23): 6727-32.

Scheible, W. R., B. Fry, A. Kochevenko, D. Schindelasch, L. Zimmerli, S. Somerville, R. Loria and C. R. Somerville (2003). An Arabidopsis mutant resistant to thaxtomin A, a cellulose synthesis inhibitor from *Streptomyces* species. *Plant Cell* **15**(8): 1781-1794.

Schlatter, D., A. Fubuh, K. Xiao, D. Hernandez, S. Hobbie and L. Kinkel (2009). Resource Amendments Influence Density and Competitive Phenotypes of *Streptomyces* in Soil. *Microbial Ecology* **57**(3): 413-420.

Schneider, A. and M. A. Marahiel (1998). Genetic evidence for a role of thioesterase domains, integrated in or associated with peptide synthetases, in non-ribosomal peptide biosynthesis in *Bacillus subtilis*. *Arch Microbiol* **169**(5): 404-10.

Schofield, C. J., J. E. Baldwin, M. F. Byford, I. Clifton, J. Hajdu, C. Hensgens and P. Roach (1997). Proteins of the penicillin biosynthesis pathway. *Curr Opin Struct Biol* **7**(6): 857-64.

Schwecke, T., J. F. Aparicio, I. Molnar, A. Konig, L. E. Khaw, S. F. Haydock, M. Oliynyk, P. Caffrey, J. Cortes, J. B. Lester and et al. (1995). The biosynthetic gene cluster for the polyketide immunosuppressant rapamycin. *Proc Natl Acad Sci U S A* **92**(17): 7839-43.

Scrutton, N. S., A. Berry and R. N. Perham (1990). Redesign of the coenzyme specificity of a dehydrogenase by protein engineering. *Nature* **343**(6253): 38-43.

Seipke, R. F. and R. Loria (2008). *Streptomyces scabies* 87-22 possesses a functional tomatinase. *J Bacteriol* **190**(23): 7684-92.

Seipke, R. F. and R. Loria (2009). Hopanoids are not essential for growth of *Streptomyces scabies* 87-22. *J Bacteriol*.

Serino, L., C. Reimmann, H. Baur, M. Beyeler, P. Visca and D. Haas (1995). Structural genes for salicylate biosynthesis from chorismate in *Pseudomonas aeruginosa*. *Mol Gen Genet* **249**(2): 217-28.

Serre, L., E. C. Verbree, Z. Dauter, A. R. Stuitje and Z. S. Derewenda (1995). The *Escherichia coli* malonyl-CoA:acyl carrier protein transacylase at 1.5-A

resolution. Crystal structure of a fatty acid synthase component. *J Biol Chem* **270**(22): 12961-4.

Shen, P. and H. V. Huang (1986). Homologous recombination in *Escherichia coli*: dependence on substrate length and homology. *Genetics* **112**(3): 441-57.

Sieber, S. A. and M. A. Marahiel (2003). Learning from nature's drug factories: nonribosomal synthesis of macrocyclic peptides. *J Bacteriol* **185**(24): 7036-43.

Silakowski, B., B. Kunze, G. Nordsiek, H. Blocker, G. Hofle and R. Muller (2000). The myxochelin iron transport regulon of the myxobacterium *Stigmatella aurantiaca* Sg a15. *Eur J Biochem* **267**(21): 6476-85.

Silakowski, B., G. Nordsiek, B. Kunze, H. Blocker and R. Muller (2001). Novel features in a combined polyketide synthase/non-ribosomal peptide synthetase: the myxalamid biosynthetic gene cluster of the myxobacterium *Stigmatella aurantiaca* Sga15. *Chem Biol* **8**(1): 59-69.

Silby, M., A. Cerdeno-Tarraga, G. Vernikos, S. Giddens, R. Jackson, G. Preston, X.-X. Zhang, C. Moon, S. Gehrig, S. Godfrey, C. Knight, J. Malone, Z. Robinson, A. Spiers, S. Harris, G. Challis, A. Yaxley, D. Harris, K. Seeger, L. Murphy, S. Rutter, R. Squares, M. Quail, E. Saunders, K. Mavromatis, T. Brettin, S. Bentley, J. Hothersall, E. Stephens, C. Thomas, J. Parkhill, S. Levy, P. Rainey and N. Thomson (2009). Genomic and genetic analyses of diversity and plant interactions of *Pseudomonas fluorescens*. *Genome Biol* **10**(5): R51.

Sim, S. H., Y. Yu, C. H. Lin, R. K. M. Karuturi, V. Wuthiekanun, A. Tuanyok, H. H. Chua, C. Ong, S. S. Paramalingam, G. Tan, L. Tang, G. Lau, E. E. Ooi, D. Woods, E. Feil, S. J. Peacock and P. Tan (2008). The Core and Accessory Genomes of *Burkholderia pseudomallei*: Implications for Human Melioidosis. *PLoS Pathog* **4**(10): e1000178.

Smirnova, A. V., L. Wang, B. Rohde, I. Budde, H. Weingart and M. S. Ullrich (2002). Control of temperature-responsive synthesis of the phytotoxin coronatine in *Pseudomonas syringae* by the unconventional two-component system CorRPS. *J Mol Microbiol Biotechnol* **4**(3): 191-6.

Song, L., F. Barona-Gomez, C. Corre, L. Xiang, D. W. Udwary, M. B. Austin, J. P. Noel, B. S. Moore and G. L. Challis (2006). Type III polyketide synthase beta-ketoacyl-ACP starter unit and ethylmalonyl-CoA extender unit selectivity discovered by Streptomyces coelicolor genome mining. *J Am Chem Soc* **128**(46): 14754-5.

Sonnhammer, E. L. L. and R. Durbin (1995). A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis (Reprinted from Gene Combis, vol 167, pg GC1-GC10, 1996). *Gene* **167**(1-2): Gc1-Gc10.

St-Onge, R., C. Goyer, R. Coffin and M. Filion (2008). Genetic diversity of *Streptomyces* spp. causing common scab of potato in eastern Canada. *Systematic and applied microbiology* **31** (6-8): 474-484.

Stachelhaus, T., A. Huser and M. A. Marahiel (1996). Biochemical characterization of peptidyl carrier protein (PCP), the thiolation domain of multifunctional peptide synthetases. *Chem Biol* **3**(11): 913-21.

Stachelhaus, T. and M. A. Marahiel (1995). Modular structure of peptide synthetases revealed by dissection of the multifunctional enzyme GrsA. *J Biol Chem* **270**(11): 6163-9.

Stachelhaus, T., H. D. Mootz, V. Bergendahl and M. A. Marahiel (1998). Peptide bond formation in nonribosomal peptide biosynthesis. Catalytic role of the condensation domain. *J Biol Chem* **273**(35): 22773-81.

Stachelhaus, T., H. D. Mootz and M. A. Marahiel (1999). The specificity-conferring code of adenylation domains in nonribosomal peptide synthetases. *Chem Biol* **6**(8): 493-505.

Staunton, J. and K. J. Weissman (2001). Polyketide biosynthesis: a millennium review. *Natural Product Reports* **18**(4): 380-416.

Stegmann, E., C. Rausch, S. Stockert, D. Burkert and W. Wohlleben (2006). The small MbtH-like protein encoded by an internal gene of the balhimycin biosynthetic gene cluster is not required for glycopeptide production. *FEMS Microbiol Lett* **262**(1): 85-92.

Strohl, W. R. (1992). Compilation and analysis of DNA sequences associated with apparent streptomycete promoters. *Nucleic Acids Res* **20**(5): 961-74.

Sweigard, J. A., F. G. Chumley and B. Valent (1992). Cloning and analysis of CUT1, a cutinase gene from *Magnaporthe grisea*. *Mol Gen Genet* **232**(2): 174-82.

Takano, E., M. Tao, F. Long, M. J. Bibb, L. Wang, W. Li, M. J. Buttner, M. J. Bibb, Z. X. Deng and K. F. Chater (2003). A rare leucine codon in *adpA* is implicated in the morphological defect of *bldA* mutants of *Streptomyces coelicolor*. *Mol Microbiol* **50**(2): 475-86.

Takeuchi, T., H. Sawada, F. Tanaka and I. Matsuda (1996). Phylogenetic analysis of *Streptomyces* spp. causing potato scab based on 16S rRNA sequences. *Int J Syst Bacteriol* **46**(2): 476-9.

Tanovic, A., S. A. Samel, L. O. Essen and M. A. Marahiel (2008). Crystal structure of the termination module of a nonribosomal peptide synthetase. *Science* **321**(5889): 659-63.

Tesch, C., K. Nikoleit, V. Gnau, F. Gotz and C. Bormann (1996). Biochemical and molecular characterization of the extracellular esterase from *Streptomyces diastatochromogenes*. *J Bacteriol* **178**(7): 1858-65.

Thompson, J. D., T. J. Gibson and D. G. Higgins (2002). Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* **Chapter 2**: Unit 2 3.

Thompson, J. D., D. G. Higgins and T. J. Gibson (1994). Improved sensitivity of profile searches through the use of sequence weights and gap excision. *Comput Appl Biosci* **10**(1): 19-29.

Tolba, S., S. Egan, D. Kallifidas and E. M. H. Wellington (2002). Distribution of streptomycin resistance and biosynthesis genes in streptomycetes recovered from different soil sites. *Fems Microbiology Ecology* **42**(2): 269-276.

Toussaint, A. and C. Merlin (2002). Mobile elements as a combination of functional modules. *Plasmid* **47**(1): 26-35.

Truper, H. G. and L. DeClari (1997). Taxonomic note: Necessary correction of specific epithets formed as substantives (nouns) "in apposition". *International Journal of Systematic Bacteriology* **47**(3): 908-909.

Tsai, S. C., H. Lu, D. E. Cane, C. Khosla and R. M. Stroud (2002). Insights into channel architecture and substrate specificity from crystal structures of two macrocycle-forming thioesterases of modular polyketide synthases. *Biochemistry* **41**(42): 12598-606.

Tunca, S., C. Barreiro, A. Sola-Landa, J. J. Coque and J. F. Martin (2007). Transcriptional regulation of the desferrioxamine gene cluster of *Streptomyces coelicolor* is mediated by binding of DmdR1 to an iron box in the promoter of the *desA* gene. *Febs J* **274**(4): 1110-22.

Ullrich, M. and C. L. Bender (1994). The biosynthetic gene cluster for coronamic acid, an ethylcyclopropyl amino acid, contains genes homologous to amino acid-activating enzymes and thioesterases. *J Bacteriol* **176**(24): 7574-86.

Uniprot, c. (2007). The Universal Protein Resource (UniProt). *Nucleic Acids Res* **35**(Database issue): D193-7.

Uppalapati, S. R., Y. Ishiga, T. Wangdi, B. N. Kunkel, A. Anand, K. S. Mysore and C. L. Bender (2007). The phytotoxin coronatine contributes to pathogen fitness and is required for suppression of salicylic acid accumulation in tomato inoculated with *Pseudomonas syringae* pv. tomato DC3000. *Mol Plant Microbe Interact* **20**(8): 955-65.

van Wezel, G. P., E. Vijgenboom and L. Bosch (1991). A comparative study of the ribosomal RNA operons of *Streptomyces coelicolor* A3(2) and sequence analysis of *rrnA*. *Nucleic Acids Res* **19**(16): 4399-403.

Ventura, M., C. Canchaya, A. Tauch, G. Chandra, G. F. Fitzgerald, K. F. Chater and D. van Sinderen (2007). Genomics of Actinobacteria: Tracing the evolutionary history of an ancient phylum. *Microbiology and Molecular Biology Reviews* **71**: 495-+.

Vernikos, G. S. and J. Parkhill (2006). Interpolated variable order motifs for identification of horizontally acquired DNA: revisiting the *Salmonella* pathogenicity islands. *Bioinformatics* **22**(18): 2196-203.

Vining, L. C. (1990). Functions of secondary metabolites. *Annual Review of Microbiology* **44**: 395-427.

Wach, M. J., J. A. Kers, S. B. Krasnoff, R. Loria and D. M. Gibson (2005). Nitric oxide synthase inhibitors and nitric oxide donors modulate the biosynthesis of thaxtomin A, a nitrated phytotoxin produced by Streptomyces spp. *Nitric Oxide-Biology and Chemistry* **12**(1): 46-53.

Wach, M. J., S. B. Krasnoff, R. Loria and D. M. Gibson (2007). Effect of carbohydrates on the production of thaxtomin A by Streptomyces acidiscabies. *Archives of Microbiology* **188**(1): 81-88.

Walker, J. B. (1979). On the development of enzymic pathways for the biosynthesis of aminocyclitol antibiotics and other idiolites. *Folia Microbiol (Praha)* **24**(3): 286-91.

Walker, J. B. (1988). Ontogeny of biosynthesis of streptomycin and related idiolites from myoinositol as a model differentiation system. *Faseb Journal* **2**(4): A582-A582.

Wang, J., A. Mushegian, S. Lory and S. Jin (1996). Large-scale isolation of candidate virulence genes of *Pseudomonas aeruginosa* by in vivo selection. *Proc Natl Acad Sci U S A* **93**(19): 10434-9.

Wang, L.-y., S.-t. Li and Y. Li (2003a). Identification and characterization of a new exopolysaccharide biosynthesis gene cluster from *Streptomyces*. *Fems Microbiology Letters* **220**(1): 21-27.

Wang, L.-y., S.-t. Li and Y. Li (2003b). Isolation and sequencing of glycosyltransferase gene and UDP-glucose dehydrogenase gene that are located on a gene cluster involved in a new exopolysaccharide biosynthesis in *Streptomyces*. *DNA Seq* **14**(2): 141-5.

Wang, X., F. AlarcÓn-Chaidez, A. PeÑaloza-VÁzquez and C. L. Bender (2002). Differential regulation of coronatine biosynthesis in *Pseudomonas syringae* pv. tomato DC3000 and *P. syringae* pv. *glycinea* PG4180. *Physiological and Molecular Plant Pathology* **60**(3): 111-120.

Washietl, S. and I. L. Hofacker (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* **342**(1): 19-30.

Washietl, S., I. L. Hofacker and P. F. Stadler (2005). Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**(7): 2454-9.

Watve, M. G., R. Tickoo, M. M. Jog and B. D. Bhole (2001). How many antibiotics are produced by the genus *Streptomyces*? *Arch Microbiol* **176**(5): 386-90.

Weber, T., R. Baumgartner, C. Renner, M. A. Marahiel and T. A. Holak (2000). Solution structure of PCP, a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure* **8**(4): 407-18.

Wei, Y., J. L. Schottel, U. Derewenda, L. Swenson, S. Patkar and Z. S. Derewenda (1995). A novel variant of the catalytic triad in the *Streptomyces scabies* esterase. *Nat Struct Biol* **2**(3): 218-23.

Weingart, H., S. Stubner, A. Schenk and M. S. Ullrich (2004). Impact of temperature on *in planta* expression of genes involved in synthesis of the *Pseudomonas syringae* phytotoxin coronatine. *Mol Plant Microbe Interact* **17**(10): 1095-102.

Weissman, K. J. (2009). Introduction to polyketide biosynthesis. In *Complex Enzymes in Microbial Natural Product Biosynthesis, Part B: Polyketides, Aminocoumarins and Carbohydrates*. San Diego, Elsevier Academic Press Inc. **459:** 3-16.

Wellington, E. M. H. and I. K. Toth (1994). Actinomycetes. In *Methods of soil analysis, part 2: microbiological and biochemical properties.* S. H. Mickleson and J. M. Bigham. Madison, Soil Science Society of America**:** 269-290.

Wennerhold, J. and M. Bott (2006). The DtxR regulon of *Corynebacterium glutamicum*. *Journal of Bacteriology* **188**(8): 2907-2918.

Wiener, P., S. Egan, A. S. Huddleston and E. M. Wellington (1998). Evidence for transfer of antibiotic-resistance genes in soil populations of streptomycetes. *Mol Ecol* **7**(9): 1205-16.

Wiggins, B. E. and L. L. Kinkel (2005). Green manures and crop sequences influence potato diseases and pathogen inhibitory activity of indigenous streptomycetes. *Phytopathology* **95**(2): 178-185.

Willetts, N. S., C. Crowther and B. W. Holloway (1981). The Insertion-Sequence Is21 of R68.45 and the Molecular-Basis for Mobilization of the Bacterial Chromosome. *Plasmid* **6**(1): 30-52.

Williams, S. T., M. Goodfellow and G. Alderson (1989). Genus *Streptomyces* Waksman and Henrici 1943. In *Bergey's manual of systematic bacteriology*. S. T. Williams, M. E. Sharpe and J. G. Holt. Baltimore, MD, Williams and Wilkins. **4 Actinomycetes:** 2452–2492.

Williams, S. T., M. Goodfellow, G. Alderson, E. M. Wellington, P. H. Sneath and M. J. Sackin (1983). Numerical classification of Streptomyces and related genera. *J Gen Microbiol* **129**(6): 1743-813.

Wilson, C. R., L. M. Ransom and B. M. Pemberton (1999). The Relative Importance of Seed-borne Inoculum to Common Scab Disease of Potato and the Efficacy of Seed Tuber and Soil Treatments for Disease Control *Journal of Phytopathology* **147**(1): 13-18.

Wisedchaisri, G., R. K. Holmes and W. G. J. Hol (2004). Crystal structure of an IdeR-DNA complex reveals a conformational change in activated IdeR for base-specific interactions. *Journal of Molecular Biology* **342**(4): 1155-1169.

Woese, C. R., O. Kandler and M. L. Wheelis (1990). Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc Natl Acad Sci U S A* **87**(12): 4576-9.

Wolpert, M., B. Gust, B. Kammerer and L. Heide (2007). Effects of deletions of *mbtH*-like genes on clorobiocin biosynthesis in *Streptomyces coelicolor*. *Microbiology* **153**(Pt 5): 1413-23.

Wright, F. and M. J. Bibb (1992). Codon usage in the G+C-rich *Streptomyces* genome. *Gene* **113**(1): 55-65.

Wu, J., T. J. Zaleski, C. Valenzano, C. Khosla and D. E. Cane (2005). Polyketide double bond biosynthesis. Mechanistic analysis of the dehydratase-containing module 2 of the picromycin/methymycin polyketide synthase. *J Am Chem Soc* **127**(49): 17393-404.

Yamanaka, K., H. Oikawa, H. O. Ogawa, K. Hosono, F. Shinmachi, H. Takano, S. Sakuda, T. Beppu and K. Ueda (2005). Desferrioxamine E produced by Streptomyces griseus stimulates growth and development of *Streptomyces tanashiensis*. *Microbiology* **151**(Pt 9): 2899-905.

Yang, C. C., C. H. Huang, C. Y. Li, Y. G. Tsay, S. C. Lee and C. W. Chen (2002). The terminal proteins of linear *Streptomyces* chromosomes and plasmids: a novel class of replication priming proteins. *Mol Microbiol* **43**(2): 297-305.

Yeats, C., S. Bentley and A. Bateman (2003). New knowledge from old: in silico discovery of novel protein domains in *Streptomyces coelicolor*. *BMC Microbiol* **3**: 3.

Yellaboina, S., S. Ranjan, P. Chakhaiyar, S. E. Hasnain and A. Ranjan (2004a). Prediction of DtxR regulon: Identification of binding sites and operons controlled by Diphtheria toxin repressor in *Corynebacterium diphtheriae*. *Bmc Microbiology* **4**: -.

Yellaboina, S., J. Seshadri, M. S. Kumar and A. Ranjan (2004b). PredictRegulon: a web server for the prediction of the regulatory protein binding sites and operons in prokaryote genomes. *Nucleic Acids Res* **32**(Web Server issue): W318-20.

Yin, X. and T. M. Zabriskie (2006). The enduracidin biosynthetic gene cluster from *Streptomyces fungicidicus*. *Microbiology* **152**(Pt 10): 2969-83.

Youard, Z. A., G. L. Mislin, P. A. Majcherczyk, I. J. Schalk and C. Reimmann (2007). *Pseudomonas fluorescens* CHA0 produces enantio-pyochelin, the optical antipode of the *Pseudomonas aeruginosa* siderophore pyochelin. *J Biol Chem* **282**(49): 35546-53.

Young, J. M. and J. P. Euzeby (2008). Proposed revision of Rule 33c to perpetuate the citation of revived names. *Int J Syst Evol Microbiol* **58**(Pt 10): 2468-9.

Zaitlin, B. and S. B. Watson (2006). Actinomycetes in relation to taste and odour in drinking water: Myths, tenets and truths. *Water Research* **40**(9): 1741-1753.

Zawadzka, A. M., R. J. Abergel, R. Nichiporuk, U. N. Andersen and K. N. Raymond (2009). Siderophore-Mediated Iron Acquisition Systems in Bacillus cereus: Identification of Receptors for Anthrax Virulence-Associated Petrobactin (dagger) (,) (double dagger). *Biochemistry*.

Zdobnov, E. M. and R. Apweiler (2001). InterProScan--an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17**(9): 847-8.

Zerikly, M. and G. L. Challis (2009). Strategies for the discovery of new natural products by genome mining. *Chembiochem* **10**(4): 625-33.

Zhang, R. and C. T. Zhang (2004). A systematic method to identify genomic islands and its applications in analyzing the genomes of *Corynebacterium glutamicum* and *Vibrio vulnificus* CMCP6 chromosome I. *Bioinformatics* **20**(5): 612-22.

Zhou, M., J. Boekhorst, C. Francke and R. J. Siezen (2008). LocateP: genome-scale subcellular-location predictor for bacterial proteins. *BMC Bioinformatics* **9**: 173.

Zhu, H. Q., G. Q. Hu, Z. Q. Ouyang, J. Wang and Z. S. She (2004). Accuracy improvement for identifying translation initiation sites in microbial genomes. *Bioinformatics* **20**(18): 3308-17.

Zirkle, R., T. A. Black, J. Gorlach, J. M. Ligon and I. Molnar (2004). Analysis of a 108-kb region of the *Saccharopolyspora spinosa* genome covering the obscurin polyketide synthase locus. *DNA Seq* **15**(2): 123-34.

Zwahlen, J., S. Kolappan, R. Zhou, C. Kisker and P. J. Tonge (2007). Structure and mechanism of MbtI, the salicylate synthase from *Mycobacterium tuberculosis*. *Biochemistry* **46**(4): 954-64.

Appendices

A Published work: **Morningstar** *et al.* 2006 (book chapter; submitted manuscript attached.)

B Published work: **Silby** *et al.* 2009 (research paper; pdf attached.)

C *On attached digital medium,* DNA and annotation files of *S. scabies* 87-22 genome for viewing in `Artemis`: `SCAB.dna`, `SCAB.tab`

D *On attached digital medium,* colour and class scheme used by annotators at WTSI to classify predicted coding sequences, `colour_and_class_scheme.txt`

E *On attached digital medium,* additional files for use with `Artemis`: `concanamycin_cluster.tab` and `concanamycin_cluster.dna` subsections of the main annotation file with in-depth annotation details for biosynthetic domains.

F *On attached digital medium,* complete list of TTA codons identified, tab separated text file `scabies_tta.txt`

G *On attached digital medium,* `SCAB85471.xlsx` Microsoft Excel 2007 spreadsheet file showing working for domains in cluster predicted to encode peptide siderophore.

EVOLVING GENE CLUSTERS IN SOIL BACTERIA

ALICE MORNINGSTAR, WILLIAM H. GAZE, SAHAR TOLBA AND ELIZABETH M. H. WELLINGTON

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL, UK

INTRODUCTION

Soil is heterogeneous in nearly all respects and contains a huge diversity of microorganisms. The availability of carbon and other energy sources, mineral nutrients, and water varies considerably over space and time, as does temperature. Adaptations to nutrient poverty including oligotrophy and zymogeny (upsurge in growth when nutrients available) are common. The water films essential for microbial life in soil are discontinuous, and only clay particles have the necessary charges to hold water against the pull of gravity. Clay-coated soil particles cluster together forming aggregates, and these aggregates or clusters of aggregates with their adjacent water form the microhabitats in which bacteria function (Stotzky, 1997). The result of the discrete microhabitats in soil is that microbial population dynamics and interactions are very different from those in well-mixed substrates such as some aquatic environments. Soil is also a reservoir for herbicides, pesticides and other chemical and microbiological inputs from slurry application, all of which will have a selective impact on the indigenous bacteria.

Bacterial evolutionary histories are difficult to untangle. Different scales of evolution occur simultaneously, from events possible over a few generations (chromosomal rearrangement, gene deletion, and acquisition of genes via horizontal transfer) to the eon-scale generative evolution which creates diversity from which the novel functional genes of the future will be selected. In the age of genomics we are developing the tools to study the ecology of microbes in soil. The few metagenomic projects undertaken so far illustrate the diversity of bacteria in soil (>3000 ribotypes in a Minnesota farm soil sample) and the technical difficulties in producing overlapping sequence, five billion base pairs of sequence would be necessary to obtain the eightfold coverage traditionally targeted for draft genome assemblies, even for the single most predominant genome (Tringe *et al.*, 2005).

This chapter focuses on the evolutionary processes that are revealed by our knowledge of gene clusters studied in soil-dwelling bacteria. In soil, large communities of diverse bacteria exist where the horizontal gene pool can provide access to adaptive traits such as xenobiotic degradation, antibiotic production and resistance. The heterogeneity of the soil environment also requires flexibility allowing bacteria to adapt to constantly changing environmental conditions and respond to challenges from anthropogenic inputs. Emphasis is placed on antibiotic production and resistance as examples of clustered genes which undoubtedly play a key role in the competitiveness of bacteria in soil, while being non-essential for growth in laboratory conditions. Soil-dwelling bacterial groups such as actinobacteria, pseudomonads and bacilli produce an impressive range of antibiotics and other secondary metabolites. Streptomycetes in particular show ability to produce multiple secondary metabolites from diverse chemical classes (Hopwood, 2004; Weber *et al.*, 2003). It has been argued (Challis & Hopwood, 2003) that synergy and contingency are important in driving evolution of multiple pathways for production of secondary metabolites and give a competitive advantage to the producer. Studying the evolution of these gene clusters will improve our understanding of bacterial growth and survival in soil.

DEFINING GENE CLUSTERS

Like many other terms in biology, "cluster" is frequently used but seldom precisely defined. A useful definition includes co-ordinated regulation of a number of adjacent transcription units which may be found in both senses, strands and any frame. Collections of genes with related function can undergo reassortment to form new pathways, and are likely to have undergone horizontal transfer at some stage of their evolution. Operons are distinct from clusters, as bacterial operons have been described as one polycistronic transcriptional unit initiated at one promoter, whereas gene clusters might have more than one promoter and several transcriptional units in various senses. Singleton & Sainsbury (2001) describe prokaryotic operons as two or more genes whose expression is (1) co-ordinated from a single promoter, from which a polycistronic mRNA is produced, or (2) a common regulatory region, mRNAs "being formed by divergent transcription from different promoters". The definition of gene cluster used here is closer to the second definition.

A method has recently been proposed to search genomes and apply an analytical statistical test for clustering (Hoberman *et al.*, 2005) where searches are made for contiguous regions containing genes with homology, separated by no more than a certain number of non-homologous genes. The important contribution of this method is in the definition of clustering used. Some degree of homology is assumed to indicate functionally related genes, which would have an origin in duplication followed by divergence. Non-homologous genes situated within clusters of related genes may function as part of the cluster. The max-gap cluster analysis (Hoberman *et al.*, 2005) allows a certain number of unrelated genes to be passed over in order to identify the higher-level functional organisation which constitutes a cluster.

ORIGIN OF GENE CLUSTERS – TANDEM DUPLICATION AND DIVERGENCE

Gene clusters are formed by selection, and linkage would be expected whenever this produces a selective advantage. Variation released in recombination must be sufficiently conservative to ensure that at least some viable cells are produced. Hence we expect a linkage effect preventing recombination between factors which interact intensely because disruption of the linkage would have a viability effect (Stahl & Murray, 1966).

Duplication of gene clusters has been observed to create massive gene dosages in *E. coli* and *Salmonella typhimurium*, where short clusters have been observed to be reversibly duplicated up to a hundred times. This suggests that another mechanism of recombination other than *recA*-mediated sister chromosome exchanges is taking place (Hughes, 1998). In streptomycetes, genomic instability seems very common, with extensive chromosomal deletions and intense amplification taking place in the apparent absence of selection pressure (Birch *et al.*, 1990). It has been proposed that gene duplication acts as a "dynamic and reversible regulatory mechanism that facilitates adaptation" (Reams & Neidle, 2004) so the selective advantage of such tandem duplications as a response to variable conditions needs to be included in consideration of gene clustering in environments such as soil.

Gene duplication followed by divergence is likely to be a feature of the histories of genes within functional clusters. Bioinformatic comparison methods can test whether

pairs of genes chosen are more or less likely to be found side by side in genomes. Paralogous genes (thought to have originated from a duplication event within an organism) are found adjacent to each other across the genomes of *E. coli* and *B. subtilis*. Unrelated genes located side by side across the boundary between cistrons are used as controls in the comparison as they are assumed not to be under selection for adjacency (Janga & Moreno-Hagelsieb, 2004). The conservation of adjacency in related genes compared to those considered to be unrelated genes across the boundary of a polycistronic transcription unit appears to extend across a large number of genomes (Moreno-Hagelsieb *et al.*, 2001).

Genes in the streptomycin cluster (Fig 1) are found in all reading frames on both strands and in both senses. The retention of clusters through evolutionary history is intriguing, given that partial loss would not necessarily result in death of the organism. It is clear that some partial clusters exist (Egan *et al.*, 2001), and this will be discussed further.

Other researchers have used the tendency of related genes to be found adjacent to each other in reverse, to predict gene pairs likely to be in an operon from their adjacent occurrence across many genomes (Dandekar *et al.*, 1998; Ermolaeva *et al.*, 2001; Overbeek *et al.*, 1999; Overbeek *et al.*, 2000; Overbeek *et al.*, 2003). Although it is clear that overall genomic order is not preserved to a great extent across bacteria (Bentley *et al.*, 2002; Mushegian & Koonin, 1996), the identification of related pairs from duplication and divergence is a starting point for tracing the evolution of a gene cluster.

THE ‘SELFISH OPERON’

It has been proposed that gene clusters form through repeated horizontal gene transfer (HGT) events, between and within taxa, the so called “selfish operon” model (Lawrence & Roth, 1996). Formation through HGT would account for retention in clusters of genes that are expected to be selectively neutral or under only weak selection. It is not likely that this applies to all gene clusters, because many of the xenobiotic-degrading gene clusters that have evolved in soil bacteria result directly from selection in response to exposure and are thought to be valuable for survival and hence under selection. Genes “not essential” under laboratory conditions might

still be essential for survival in soil, and hence selectable and able to cluster via linkage effects.

A sustained attack on the selfish operon model has resulted in some interesting work testing hypotheses regarding the expected distribution of clustering. It has been found in whole genome analysis of *E. coli* that essential genes with related functions have a stronger tendency to cluster than non-essential genes (Pal & Hurst, 2004). This is of course the expected result, as linkage is expected between genes with related function. Another test examined whether horizontally transferred genes were more or less likely to be found in operons (Price *et al.*, 2005), and such investigations into clustering mechanisms have obvious relevance to clustering processes in soil bacteria.

It seems that organisation in bacterial genomes arises at operon level and above despite frequent gene rearrangements (Reams & Neidle 2004). Co-transcription appears to be insufficient to account for the structure of gene clusters in soil bacteria, with genes organised together on the genome but in both senses and different frames. Gene amplification can work alongside HGT as an evolutionary mechanism involved in gene clustering. Gene clusters are often found on mobile genetic elements (MGEs) such as plasmids, phages and genomic islands (GEIs) and can confer a new phenotypic trait allowing improved adaptation to a specific niche. However it has been argued that HGT is not a cause of operon formation but instead promotes the prevalence of pre-existing operons (Price *et al.*, 2005).

SECONDARY METABOLISM AND ANTIBIOTIC GENE CLUSTERS

Antibiotics are synthesized by bacteria in soil (Anukool *et al.*, 2004) and the biosynthetic gene clusters responsible for such metabolite production are co-ordinately regulated (Chater & Bibb, 1997), have been subject to HGT (Egan *et al.*, 2001) and are discontinuous in their distribution within closely related groups of species. The gene clusters usually have a chromosomal location but several have been found on plasmids particularly linear plasmids in the case of streptomycetes (Kinashi *et al.*, 1994; Mochizuki *et al.*, 2003). Within the *Streptomycetaceae* there is an extensive diversity of secondary metabolites; some, like streptomycin, have a

defined biological activity while others such as geosmin, a volatile isoprenoid compound (responsible for the soil's 'earthy odour'), have unknown activity. Production of secondary metabolites provides a selective advantage in soil where the role of the metabolite may be as molecular weapons, repellents, attractants, 'nutriophores' (e.g. siderophores), detoxification/replacement metabolism, co-catalysts, enzyme or transport inhibitors, shaped by a long evolutionary history for highly specific interaction with their molecular targets (Piepersberg, 1993; Piepersberg, 2002).

Streptomycetes are ubiquitous soil bacteria and can be readily isolated from the majority of soils; they are filamentous, having a complex life cycle: spores germinate and form vegetative and then aerial hyphae which mature into spore chains that are subsequently released. Sequencing of the *Streptomyces coelicolor* chromosome has revealed an "unprecedented proportion of regulatory genes" thought to be involved in response to external stimuli and stresses (Bentley *et al.*, 2002). More than 20 gene clusters have been found in the *S. coelicolor* genome. Sets of duplicated genes are suggested to provide "tissue specific" responses depending on the developmental phase of the cell surrounding the nucleoid.

More than a million base pairs of genetic material at either end of the *S. coelicolor* linear chromosome can be deleted without affecting viability under lab conditions (Bentley *et al.*, 2002). These 'arm' regions of the chromosome are found to be variable and seem to contain an increasing proportion of functions related to secondary metabolism and resistance towards the ends. The core region close to the origin of replication by contrast contains most of the essential housekeeping genes such as cell cycling, energy, protein and DNA metabolism. The core has been found to be more stable, with gene order conserved across different strains and even amongst distantly related organisms such as *Mycobacterium tuberculosis* and *Corynebacterium diphtheriae* (Bentley et al., 2002).

When different strains encounter each other, formation of partial diploids as the chromosomes align followed by crossovers between chromosomes and homologous recombination can result in recombination of up to a fifth of the genome, which is a large amount of genetic material in an organism with an 8 Mb chromosome. The linear chromosome, with bound telomeres means that a single crossover can cause

exchange of genetic material. It might be assumed that this is a rare event but the formation of one crossover instead of two for an exchange is uniquely created by the linear structure of the chromosome. The more frequently occurring case where both telomeres are inherited from the same parent causes a high degree of linkage to be observed between the telomeres; hence the circular linkage map first observed by D. A. Hopwood. The trigger for crossovers and subsequent rearrangements is still obscure, but the elevated frequency of such events under the stress of mutagenic treatment and the involvement of illegitimate recombination indicate that they may be the result of an SOS-like response (Birch *et al.*, 1990), with little evidence to support a connection between mobile DNA elements and the phenomenon of genetic instability.

Recent advances in the study of recombination in *Acinetobacter* (de Vries & Wackernagel, 2002; de Vries et al., 2004) reveal one possible basis for such extraordinary feats of recombination. It is proposed that homologous recombination within an anchor region couples the donor and recipient DNA, followed by an illegitimate recombination event integrating the section. The DNA mismatch repair system in *Pseudomonas stutzeri*, another ubiquitous soil bacterium, has been implicated in the acquisition of foreign DNA from plants as well as other microbes, and seems to affect this homology-facilitated illegitimate recombination process (Meier & Wackernagel, 2005). Another group report homology independent illegitimate recombination, also in *Acinetobacter* (Reams & Neidle, 2004). Although the detail has yet to be elucidated, the evidence of such transfer processes support the conclusions from phylogenetic studies of HGT as a major factor in the evolution and adaptation of microbes (Vining, 1992).

In some gene clusters (e.g. streptomycin biosynthesis, Fig. 1) genes are found in opposing senses and different strands within the cluster. It would appear that genes are transcribed in opposite directions. When this occurs amongst closely spaced promoters, it has been suggested that the supercoiling caused by multiple polymerases affects expression of genes or operons (Opel *et al.*, 2001). So the arrangement of genes in the streptomycin biosynthesis cluster (Fig 1) could be associated with divergent transcription effects altering the expression of genes.

Transcriptional coupling of this kind is likely to serve a physiological purpose: in *E. coli*, global superhelical density of the chromosome is controlled by the energy charge of the cell, which is affected by environmental stresses and transition between growth states (Opel *et al.*, 2001). The twin-domain model (Liu & Wang, 1987) suggests that closely spaced, divergent, superhelically sensitive promoters can affect the transcriptional activity of one another by transcriptionally induced negative DNA supercoiling generated in the divergent promoter region (Rhee *et al.*, 1999).

Transcriptional interference is thought to be very widespread, where the activity of one transcription complex affects the activity of others. Studies in *E. coli* (Callen *et al.*, 2004) show effects on transcription due to the formation or other transcriptional complexes, even at distant promoters. Although the dynamics of transcriptional interference do not seem to conform to a simple model (Shearwin *et al.*, 2005; Sneppen *et al.*, 2005), perhaps consideration of multiple polymerase effects may throw light on the arrangement of genes in gene clusters.

In the streptomycin biosynthesis cluster in *S. griseus*, several pairs of genes lie in opposing senses. This would prevent polycistronic transcription, but divergent transcription might account for another level of regulation from the arrangement of the genes. However, the specific arrangement of genes in an antibiotic cluster might not affect antibiotic synthesis. Biosynthetic clusters producing chromomycin in *S. griseus* (Menendez *et al.*, 2004) and mithramycin in *S. argillaceus* (Rohr *et al.*, 1999), despite making very similar chemical structures and having nearly identical genes, are ordered quite differently (O'Connor, 2004).

SELECTIVE PRESSURES FAVOURING GENE CLUSTERING

Firn & Jones (2000) put forward the view that not every product of secondary metabolism has to be bioactive in order for the ability to create these substances to be retained by selection. The selective advantage of the rare bioactive molecule is clear, and it seems reasonable to suppose that natural selection favours organisms able to generate and maintain chemical novelties at the least cost. By analogy with the animal immune system, possession of the mechanism for generating metabolites is crucial, and the waste resulting from the many non-active molecules generated, like the many non-useful antibodies, can be tolerated.

Alteration of substrate specificity is likely to be a favourable mutation in an enzyme involved in metabolism. In primary metabolism, enzymes seem to have very precise substrate specificity, probably due to the high cost of inefficient processing in the cell's most essential functions. Retention in secondary metabolism of enzymes with broad substrate specificity would be expected, given the different selection pressures on such products (Firn & Jones, 2000). Enzymes with broad specificity will have a greater chance to generate novel products: a single mutation in spearmint produced novel peppermint flavours (Croteau, 1991). Enzymes further down the monoterpene production pathway were apparently able to act on the modified substrates to produce the novel products, and perhaps this broad specificity would not be expected in primary metabolism.

Vining (1992) focuses on the survival value of secondary metabolites for the organism. He suggests that secondary metabolism is found when the organism exists in a competitive environment. The soil environment is likely to fit these criteria. Different kinds of organisms, plants, fungi, arthropods, as well as many kinds of microbes are competing for the nutrient resources available in soil. In harsher environments where primarily anaerobes or chemolithotrophs are found, resources are so scarce that few types of organism are found, competition is less, and metabolism is more streamlined with fewer secondary metabolic products. Vining (1992) points out that many of the bioactive metabolites produced have specificity for a particular kind of competitor, whether a grazer of mycelium (e.g. arthropods targeted by avermectin produced by *S. avermitilis* (Hotson, 1982) or a more direct competitor for nutrition (e.g. siderophore pseudobactin 358, produced by *Pseudomonas putida* in the rhizosphere, which seems to limit growth of other bacteria and fungi by sequestering available iron (de Weger *et al.*, 1988)).

Genome comparison between *S. coelicolor* and *S. avermitilis* has been used to provide a model for the evolution of these large linear chromosomes where the central 'core' region is derived from a common actinobacterial ancestor, and contains mostly essential and housekeeping genes, while the chromosome 'arms' comprise laterally acquired contingency genes, which have been added and retained for their circumstantially advantageous qualities (Karoonuthaisiri *et al.*, 2005). The genome of *S. avermitilis* clearly supports this model (Ikeda *et al.*, 2003). Comparison

with *S. coelicolor* shows that the core regions are very similar while the arms are almost completely different and are likely to have developed independently. Ikeda *et al.*, (2003) also note that the origin of replication for the *S. avermitilis* chromosome is 776 kb away from the centre, however, it does appear to lie central to the core region so it might be the uneven size of the arms that gives the impression of asymmetry. Intriguingly, Ikeda et al. also note that although most of the secondary metabolism clusters are located in the arms, the few that lie in the core tend to be common to several streptomycete species, suggesting that they were fixed in the chromosome at an evolutionarily early stage. The streptomycete linear chromosome possibly provides a distinct advantage for horizontal transfer. Rare recombinations near one end of the chromosome could result in exchange of the unstable arm region between strains, as depicted in Fig. 2.

ORIGIN OF GENE CLUSTERS – HGT

Vining (1992) has compared nucleotide sequences from polyketide and phenazine antibiotic synthesis pathways with related genes in primary pathways. Where sequence similarity exists to primary pathway genes in distantly related organisms, it seems likely that the host in which the secondary metabolite gene was built over deep time has been identified. Vining suggests from his study of fatty acid synthase and polyketide synthase (PKS) pathways that certain core secondary metabolite pathways have "arisen early in evolution and been maintained in reservoir organisms" from which they have been donated via gene transfer. If this is the case, he suggests we would expect to see an ancient core component in widely distributed secondary metabolic pathways, with more recently evolved terminal and specificity-conferring reactions "grafted on" to the ancient core of the pathway. The early steps in which intermediates of little selective value are generated might have arisen only once, but horizontal transfer allows their acquisition once they are linked in a pathway which provides a selective advantage.

The analogy between polyketide and long-chain fatty acid biosynthesis has been extended by studies that have demonstrated similarity between the products of several PKS genes (Hopwood & Sherman, 1990; Katz & Donadio, 1993) and there fatty acid synthase congeners in *E. coli* (Magnuson *et al.*, 1993; Vanden Boom & Cronan, 1989) yeast (Schweizer *et al.*, 1986) and mammals (Witkowski *et al.*, 1991).

Examples have been found of PKSs resembling each of the classical classes of fatty acid synthases (FASs). The PKSs are classified in to three types of which the Type I and II are found mainly in bacteria and fungi. It is believed that the Type I and II PKSs share the same evolutionary origin but their sequences are too far diverged for significant DNA hybridisation (McCarthy *et al.*, 1983).

PKSs are multifunctional enzymes which carry out repeated rounds of carbon chain building by condensation of carboxylic acids to form a polyketide chain. Metsa-Ketela *et al.*, (1999) studied the molecular diversity of a portion of the ketosynthase gene, KSα, responsible for the building of the carbon chain (in the aromatic polyketide pathway) in a wide range of soil isolates of *Streptomyces* species. Comparison of KSα gene sequence phylogeny with that obtained from analysis of the γ-variable 16S rRNA gene indicated extensive HGT. There was no evidence of correlation between the two gene phylogenies and very similar KSα sequences were recovered in distantly related species while high KSα sequence diversity was evident within species. Gene clusters for glycopeptide antibiotic biosynthesis showed strong evidence for mosaicism where Donadio *et al.*, (1991) reported on the comparison of five gene clusters involved in biosynthesis of structurally related antibiotics which shared a similar mechanism of action on the bacterial cell wall. Conserved synteny was observed between clusters and preceding intergenic regions were often conserved. The clusters consisted of distinct gene cassettes encoding enzymes for sub-pathways which expanded by gene acquisition adding specific tailoring steps, regulatory and resistance genes. Extensive conserved synteny has also been observed in comparative analysis of gene clusters (PKSs, post PK modifications and regulatory genes) involved in polyene antibiotic biosynthesis (Aparicio *et al.*, 2003). The clusters for nystatin, primaricin, amphotericin and candicidin were compared and evidence for gene duplications was noted and thought to have played a recent role in the evolution of the gene clusters. Many polyenes show potent antifungal activity and Aparicio *et al.*, (2003) hypothesised that the large polyene-producing PKS systems could have evolved from the smaller PKSs that assemble macrolides inhibiting prokaryotic ribsomes.

THE STREPTOMYCIN CLUSTER - ECOLOGICAL CONTEXT

The streptomycin biosynthetic gene cluster is most studied in its ecological context. Streptomycin use in horticulture has never been widespread in Europe, and has been banned; although it was used to treat fireblight (*Erwinia*) infection in orchards, and was also occasionally used in veterinary medicine. Streptomycin is one of the best studied antibiotics, several studies of the ecology of producers has elucidated the evolutionary context in which the genes are evolving (Egan *et al.*, 2001; Tolba *et al.*, 2002). Density and spatial clustering of the producing strain appears to have an impact on antibiotic production (Wiener, 2000), which also seems to be highly variable (Davelos *et al.*, 2004a; Davelos *et al.*, 2004b). In genetic analysis of *Streptomyces* strains implicated in potato scab disease, pathogenicity and phenotype do not seem to correlate well with phylogeny inferred by 16S rRNA sequence (Bramwell *et al.*, 1998). This complicates the task of inferring past, present, and future evolutionary processes. Strains from sites where streptomycin has been applied (as plantomycin) compared to those from untreated sites or sites where sewage sludge has been applied (Tolba *et al.*, 2002), showed that the level of resistance to streptomycin in recovered stains was similar regardless of treatment. Screening of these soil isolates from various European locations (Tolba *et al.*, 2002) for *strA* (thought to have a role in streptomycin resistance), and *strB1* (the neighbouring gene in *S. griseus* which is involved in biosynthesis) revealed increased prevalence of both genes in the streptomycin-treated soil.

The presence of streptomycin biosynthesis genes in non-producing strains is interesting. Antibiotic production by streptomycetes has been shown to have a role in preventing the invasion of sensitive competitors in lab conditions, as compared to either mediation between co-inoculated *S. griseus* and *B. subtilis* or invasion of *S. griseus* into an established population of *B. subtilis* (Wiener, 1996). When production as well as resistance is conferred by the gene cluster, it is reasonable to suggest that production has a role in defending localised nutrients from sensitive strains. When streptomycin is present in the environment without direct metabolic cost (i.e., from anthropogenic application) the cost of producing streptomycin is less worthwhile, as fewer sensitive strains can be assumed to be present. Acquisition of partial clusters may allow resistance to streptomycin (e.g. via modification of the streptomycin molecule), hence we could expect those genes to be retained, unless a

ribosome mutation also conferring resistance (target modification) was found in the organism.

Although prevalence of *strA* and *strB1* genes in DNA extracted from soil at the plantomycin treated site was high, few producers were isolated. This could indicate that some strains had acquired the genes but not the regulatory apparatus necessary for production. Since the usefulness of production is reduced by anthropogenic application, it might be expected that some strains would lose production capability, since the selective advantage would have disappeared. However, if we expect that evolution favours organisms able to produce non-essential metabolites at lower cost (Firn & Jones, 2000; Firn & Jones, 2003) we might expect to recover producers, non-producers with some production genes, and strains with ribosome mutations conferring resistance without the presence of production genes alongside each other. The fitness cost of maintaining biosynthetic genes would not be sufficient to confer an advantage to strains which had lost them. The possibilities for selection of production and resistance are summarised in Fig. 3 where in case 1 streptomycin is produced, and in case 2 streptomycin produced but anthropogenic streptomycin also encountered. Case 3 indicates that the loss of production capacity reduces metabolic costs, and case 4 the acquisition of "target mutation" in ribosome sequence confers resistance without the cost of maintaining any biosynthetic genes. The study (Tolba *et al.*, 2002) also showed that streptomycetes genetically similar to *S. coelicolor* had *strA* and *strB1* genes with high homology to those found in *S. griseus* type strains, indicative of horizontal transfer. Other incidences have been found of transfer between streptomycetes, e.g. from Brazilian soils (Huddleston *et al.*, 1997). This would be similar to case (3) in Fig. 3 above. MGEs have not as yet been implicated in the mobility of the *str* cluster although linear plasmids and GEIs are thought to play a role in mobilising functional antibiotic gene clusters.

Future work might investigate the suggestion (Vining, 1992) that the biosynthetic cluster had assembled in one strain followed by transfer to others. Comparison of sugar handling genes in the pathway might quantify mutation rates of genes involved in polyketide sugar unit synthesis and nucleotide sugar metabolism compared to those in streptomycin biosynthesis clusters. This kind of study might identify the

sources of the genes which have diverged to perform streptomycin biosynthesis by which strains have the highest homology between those and sugar handling genes.

EVOLUTION OF XENOBIOTIC-DEGRADING GENE CLUSTERS

The ability to degrade xenobiotic compounds added to soil has evolved in a wide range of bacteria and resulted in the development of gene clusters. Various mechanisms of gene acquisition, involving both MGEs and HGT have been implicated to allow catabolism of anthropogenic compounds. It is also evident that pathways for degradation of man-made pesticides, herbicides and other xenobiotics have evolved from pre-existing clusters of genes required for degradation of naturally occurring chemically similar metabolic products.

The possible evolutionary events involved in development of catabolic pathways for combating potential toxicity and novelty of xenobiotics have been extensively reviewed (Copley, 2000; Johnson & Spain, 2003; Reams & Neidle, 2004; Springael & Top, 2004). Studies of these catabolic pathways has revealed much about how bacteria evolve and adapt to new substrates, a striking example of this is the gene cluster for catabolism of 2,4-dinitrotoluene (2,4-DNT) in *Burkholderia cepacia* (Johnson & Spain, 2003). The 2,4-DNT gene cluster is comprised of three modules, each with specific dioxygenases, their phylogeny relates to aromatic acid, benzenoid and naphthalene dioxygenases respectively. Gene recruitment resulting in a cluster of genes from different, existing pathways was evident in the evolution of the pathway for pentachlorophenol (PCP) degradation (Copley, 2000). Soil bacteria have evolved the ability to degrade this pesticide since its introduction in the early part of the 20th Century. The *pcp* genes have been found clustered on two chromosomal fragments in *Sphingobium chlorophenolicum* (Dai & Copley, 2004) and there is evidence that these genes have been subject to HGT within bacterial communities (Tiirola *et al.*, 2002).

Dejonghe *et al.*, (2000) proved that bioaugmentation of soil could be used to enhance the rate of 2, 4-dichlorophenoxyacetic acid (2,4-D) degradation by addition of donor strains carrying the *tfd* genes on plasmids. The catabolic genes were transferred to the indigenous community via conjugation while the donor strain did not persist in the soil. Other MGEs have also been described in the dissemination of catabolic

activities in soil and include genomic islands (Dobrindt *et al.*, 2004) defined as large chromosomal regions flanked by repeat sequences, they contain integrases or transposases and may achieve integration via a tRNA gene as in the case of the *clc* element, encoding the genes for degradation of 3-chlorobenzoate (Muller *et al.*, 2003). GEIs clearly play a role in facilitating further evolution of gene clusters and subsequent dissemination.

INTEGRONS, CO-ORDINATELY REGULATED MOBILE GENE CLUSTERS

An interesting example of gene clustering is represented by a grouping of genes into structures known as integrons; these are recombination and expression systems that capture genes as part of a genetic element known as a gene cassette (Recchia & Hall, 1995). There are different types of integrons, one type sometimes referred to as "super integrons" which are host specific, chromosomally encoded and are more stable in their gene order and number of genes, and highly mobile "resistance integrons" situated on plasmids, transposons and the bacterial chromosome such as class1 integrons which carry a diverse range and number of genes (Fluit & Schmitz, 2004). These definitions of "super-intergons" and resistance-integrons are subject to debate (Hall & Stokes, 2004), but are still widely used. Most cassettes of known function found within non-species specific resistance integrons confer antibiotic or quaternary ammonium compound (QAC) resistance. Multidrug resistance (MDR) among *Enterobacteriaceae* is strongly linked with the presence of integrons (Leverstein-van Hall *et al.*, 2003). In this discussion the term integron will be used to refer to "resistance integrons" unless otherwise stated.

Integrons are examples of genetic elements that carry clusters or groupings of genes due to their ability to co-ordinately mobilise and regulate expression of cassette genes. A promoter in the 5'-conserved segment of class 1 integrons directs transcription of nearly all the cassette encoded genes (Hall & Collis, 1995), with a few exceptions such as *qacE,* which conferrs biocide resistance, possessing its own promoter (Ploy *et al.*, 1998). Class 1 integrons (which are the most widely studied class) are found with varying numbers of gene cassettes or none at all (Rosser & Young, 1999). The maintenance of empty integrons may allow integration of cassette genes if bacteria encounter conditions where cassette encoded genes confer

an advantage; as cassette genes are known to be excised and reintegrated in response to selective pressure. Class 1 integrons are often mobile, either being active transposons or being derivatives that are defective in self-transposition (Partridge *et al.*, 2002). Loss of transposase and/or resolvase genes from the 3'-conserved segment prevents self-transposition, although experimental evidence of transposition in non-active transposon integrons suggests that movement is still possible if requisite gene products are supplied in *trans*. Integrons are also often situated on plasmids which increases their already considerable potential for HGT.

Recent studies on integrons from soil total community DNA (TCDNA) have revealed that there are numerous different integron classes, with 14 new undescribed integrase genes discovered in heavy metal contaminated mine tailings (Nemergut *et al.*, 2004). A gene cassette showing high identity to a gene that codes for a step in a pathway for nitroaromatic catabolism was also discovered in DNA extracted from mine tailings, suggesting that integrons may be important in the HGT of genes other than antibiotic and biocide resistance.

The vast majority of class 1 integrons have been reported from the *Enterobacteriacea*, although they have been rarely reported from Gram-positive bacteria such as *Corynebacterium glutamicum* (Nesvera *et al.*, 1998). However, a recent study of class 1 integron prevelance in chicken litter revealed that *Corynebacterium, Staphylococcus, Aerococcus* and *Brevibacterium* sp. harboured the genetic element. These species comprised >85% of the litter community compared to <2% *Enterobacteriacea* (Nandi et al., 2004). Remarkably, "the concentration of *intI1* genes ranged from 50- to 500-fold greater than the concentration of cultivatable aerobic Gram-negative bacteria, the assumed major hosts for class 1 integrons". The chickens from which the litter was sampled were treated with a range of antibiotics, antibacterials and coccidiostats; it is thought that the use of antimicrobials in medicine and the farm environment may have resulted in an increased prevalence of integrons and the diversity of gene cassettes they carry. A study presently underway in the authors' laboratory has highlighted a possible correlation between anthropogenic activity and class 1 integron prevalence in TCDNA extracted from soils (Abdouslam, pers. comm.). A previous study suggested

a correlation between exposure to industrial biocides and the prevalence of class 1 integrons in culturable aerobes (Gaze *et al.*, 2005).

It is becoming clear that integrons are far more widely distributed in bacterial taxa and that the functional diversity of cassette genes is greater than previously thought. Coupled with the ability of cassette genes, integrons and mobile genetic elements bearing integrons to undergo HGT, the potential for these gene clusters to proliferate in soil bacteria exposed to a wide variety of selective pressures is considerable and is contributing to the evolution of antibiotic resistance in clinically important taxa.

CONCLUSIONS

Bacteria in soil must respond to constant fluxes in growth conditions and survive competition from a diverse microbial community. The ability to acquire a useful phenotypic trait, or multiple traits, encoded by a gene cluster provides the opportunity of acquiring a large number of genes in one step. The importance of the horizontal gene pool has constantly featured in explanations for the evolution of gene clusters with the development of patchwork catabolic pathways combining enzymes from two or more pathways. In addition the formation of gene cluster mosaics has resulted from the expansion of biosynthetic diversity achieved by recruitment of a cassette of genes encoding a specific part of a subpathway. Gene duplication and mutation play important roles in allowing changes in enzyme specificity and the evolution of new traits.

Evolution of antibiotic biosynthetic pathways may be achieved by continuous mixing of subsets of gene clusters and the formation of new ones. Certain core secondary metabolic pathways which had arisen early in evolution could have been maintained in reservoir organisms that served as donors for gene transfer. Widely distributed pathways such as polyketide formation with a common biochemistry may prove to have a relatively ancient core component on to which the more recent processing reactions responsible for the species specificity of the product have been recruited. Clearly the organisation of antibiotic biosynthesis genes in clusters has greatly facilitated the transfer and expansion of diversity within antibiotic-producing bacteria. Specialised MGEs have evolved to facilitate acquisition of a series of genes often coding for multiple defence mechanisms such as antibiotic resistance and

enable high levels of gene expression. In soil, anthropogenic inputs are dramatically changing the horizontal gene pool and this may have serious impacts on dissemination of certain traits such as antibiotic resistance.

ACKNOWLEDGEMENTS

We gratefully acknowledge financial support from the Natural Environment Research Council, grant NER/A/S/2000/01253 and the Biotechnology and Biological Sciences Research Council, grant 88/GM114200 plus a NERC studentship (AM).

REFERENCES

Anukool, U., Gaze, W. H. & Wellington, E. M. (2004). In situ monitoring of streptothricin production by Streptomyces rochei F20 in soil and rhizosphere. *Appl Environ Microbiol* **70**, 5222-5228.

Aparicio, J. F., Caffrey, P., Gil, J. A. & Zotchev, S. B. (2003). Polyene antibiotic biosynthesis gene clusters. *Appl Microbiol Biotechnol* **61**, 179-188.

Bentley, S. D., Chater, K. F., Cerdeno-Tarraga, A. M. & other authors (2002). Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2). *Nature* **417**, 141-147.

Birch, A., Hausler, A. & Hutter, R. (1990). Genome rearrangement and genetic instability in Streptomyces spp. *J Bacteriol* **172**, 4138-4142.

Bramwell, P. A., Wiener, P., Akkermans, A. D. & Wellington, E. M. (1998). Phenotypic, genotypic and pathogenic variation among streptomycetes implicated in common scab disease. *Lett Appl Microbiol* **27**, 255-260.

Callen, B. P., Shearwin, K. E. & Egan, J. B. (2004). Transcriptional interference between convergent promoters caused by elongation over the promoter. *Molecular Cell* **14**, 647-656.

Challis, G. L. & Hopwood, D. A. (2003). Synergy and contingency as driving forces for the evolution of multiple secondary metabolite production by Streptomyces species. *Proc Natl Acad Sci U S A* **100 Suppl 2**, 14555-14561.

Chater, K. F. & Bibb, M. J. (1997). Regulation of bacterial antibiotic production. In *Products of secondary metabolism* pp. 57-105. Edited by H. Kleinkauf & H. von Dohren. Weinheim, Germany.: VCH.

Copley, S. D. (2000). Evolution of a metabolic pathway for degradation of a toxic xenobiotic: the patchwork approach. *Trends Biochem Sci* **25**, 261-265.

Croteau, R. (1991). Metabolism of monoterpenes in mint (Mentha) species. *Planta Med* **57 (Suppl)**, 10-14.

Dai, M. & Copley, S. D. (2004). Genome shuffling improves degradation of the anthropogenic pesticide pentachlorophenol by Sphingobium chlorophenolicum ATCC 39723. *Appl Environ Microbiol* **70**, 2391-2397.

Dandekar, T., Snel, B., Huynen, M. & Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* **23**, 324-328.

Davelos, A. L., Kinkel, L. L. & Samac, D. A. (2004a). Spatial variation in frequency and intensity of antibiotic interactions among streptomycetes from prairie soil. *Applied and Environmental Microbiology* **70**, 1051-1058.

Davelos, A. L., Xiao, K., Flor, J. M. & Kinkel, L. L. (2004b). Genetic and phenotypic traits of streptomycetes used to characterize antibiotic activities of field-collected microbes. *Canadian Journal of Microbiology* **50**, 79-89.
de Vries, J. & Wackernagel, W. (2002). Integration of foreign DNA during natural transformation of Acinetobacter sp by homology-facilitated illegitimate recombination. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 2094-2099.
de Vries, J., Herzfeld, T. & Wackernagel, W. (2004). Transfer of plastid DNA from tobacco to the soil bacterium Acinetobacter sp by natural transformation. *Molecular Microbiology* **53**, 323-334.
de Weger, L. A., van Arendonk, J. J., Recourt, K., van der Hofstad, G. A., Weisbeek, P. J. & Lugtenberg, B. (1988). Siderophore-mediated uptake of Fe3+ by the plant growth-stimulating Pseudomonas putida strain WCS358 and by other rhizosphere microorganisms. *J Bacteriol* **170**, 4693-4698.
Dejonghe, W., Goris, J., El Fantroussi, S., Hofte, M., De Vos, P., Verstraete, W. & Top, E. M. (2000). Effect of dissemination of 2,4-dichlorophenoxyacetic acid (2,4-D) degradation plasmids on 2,4-D degradation and on bacterial community structure in two different soil horizons. *Appl Environ Microbiol* **66**, 3297-3304.
Dobrindt, U., Hochhut, B., Hentschel, U. & Hacker, J. (2004). Genomic islands in pathogenic and environmental microorganisms. *Nat Rev Microbiol* **2**, 414-424.
Donadio, S., Staver, M. J., McAlpine, J. B., Swanson, S. J. & Katz, L. (1991). Modular organization of genes required for complex polyketide biosynthesis. *Science* **252**, 675-679.
Egan, S., Wiener, P., Kallifidas, D. & Wellington, E. M. (2001). Phylogeny of Streptomyces species and evidence for horizontal transfer of entire and partial antibiotic gene clusters. *Antonie Van Leeuwenhoek* **79**, 127-133.
Ermolaeva, M. D., White, O. & Salzberg, S. L. (2001). Prediction of operons in microbial genomes. *Nucleic Acids Research* **29**, 1216-1221.
Firn, R. D. & Jones, C. G. (2000). The evolution of secondary metabolism - a unifying model. *Mol Microbiol* **37**, 989-994.
Firn, R. D. & Jones, C. G. (2003). Natural products - a simple model to explain chemical diversity. *Natural Product Reports* **20**, 382-391.
Fluit, A. C. & Schmitz, F. J. (2004). Resistance integrons and super-integrons. *Clin Microbiol Infect* **10**, 272-288.
Gaze, W. H., Abdouslam, N., Hawkey, P. M. & Wellington, E. M. (2005). Incidence of class 1 integrons in a quaternary ammonium compound-polluted environment. *Antimicrob Agents Chemother* **49**, 1802-1807.
Hall, R. M. & Collis, C. M. (1995). Mobile gene cassettes and integrons: capture and spread of genes by site-specific recombination. *Mol Microbiol* **15**, 593-600.
Hall, R. M. & Stokes, H. W. (2004). Integrons or super integrons? *Microbiology* **150**, 3-4.
Hoberman, R., Sankoff, D. & Durand, D. (2005). The statistical significance of max-gap clusters. *Comparative Genomics* **3388**, 55-71.
Hopwood, D. A. & Sherman, D. H. (1990). Molecular genetics of polyketides and its comparison to fatty acid biosynthesis. *Annu Rev Genet* **24**, 37-66.
Hopwood, D. A. (2004). Cracking the polyketide code. PLoS Biol 2, E35.
Hotson, I. K. (1982). The avermectins: A new family of antiparasitic agents. *J S Afr Vet Assoc* **53**, 87-90.

Huddleston, A. S., Cresswell, N., Neves, M. C., Beringer, J. E., Baumberg, S., Thomas, D. I. & Wellington, E. M. (1997). Molecular detection of streptomycin-producing streptomycetes in Brazilian soils. *Appl Environ Microbiol* **63**, 1288-1297.

Hughes, D. (1998). Impact of homologous recombination on genome organization and stability. In *Organisation of the prokaryotic genome*, pp. 109-128. Edited by R. L. Charlebois. Washington D.C.: ASM Press.

Ikeda, H., Ishikawa, J., Hanamoto, A., Shinose, M., Kikuchi, H., Shiba, T., Sakaki, Y., Hattori, M. & Omura, S. (2003). Complete genome sequence and comparative analysis of the industrial microorganism Streptomyces avermitilis. *Nat Biotechnol* **21**, 526-531.

Janga, S. C. & Moreno-Hagelsieb, G. (2004). Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Research* **32**, 5392-5397.

Johnson, G. R. & Spain, J. C. (2003). Evolution of catabolic pathways for synthetic compounds: bacterial pathways for degradation of 2,4-dinitrotoluene and nitrobenzene. *Appl Microbiol Biotechnol* **62**, 110-123.

Jung, Y. G., Kang, S. H., Hyun, C. G., Yang, Y. Y., Kang, C. M. & Suh, J. W. (2003). Isolation and characterization of bluensomycin biosynthetic genes from Streptomyces bluensis. *FEMS Microbiol Lett* **219**, 285-289.

Karoonuthaisiri, N., Weaver, D., Huang, J., Cohen, S. N. & Kao, C. M. (2005). Regional organization of gene expression in Streptomyces coelicolor. *Gene* **353**, 53-66.

Katz, L. & Donadio, S. (1993). Polyketide synthesis: prospects for hybrid antibiotics. *Annu Rev Microbiol* **47**, 875-912.

Kinashi, H., Mori, E., Hatani, A. & Nimi, O. (1994). Isolation and characterization of linear plasmids from lankacidin-producing Streptomyces species. *J Antibiot (Tokyo)* **47**, 1447-1455.

Lawrence, J. G. & Roth, J. R. (1996). Selfish operons: Horizontal transfer may drive the evolution of gene clusters. *Genetics* **143**, 1843-1860.

Leverstein-van Hall, M. A., HE, M. B., AR, T. D., Paauw, A., Fluit, A. C. & Verhoef, J. (2003). Multidrug resistance among Enterobacteriaceae is strongly associated with the presence of integrons and is independent of species or isolate origin. *J Infect Dis* **187**, 251-259.

Liu, L. F. & Wang, J. C. (1987). Supercoiling of the DNA template during transcription. *Proc Natl Acad Sci U S A* **84**, 7024-7027.

Magnuson, K., Jackowski, S., Rock, C. O. & Cronan, J. E., Jr. (1993). Regulation of fatty acid biosynthesis in Escherichia coli. *Microbiol Rev* **57**, 522-542.

McCarthy, A. D., Goldring, J. P. & Hardie, D. G. (1983). Evidence that the multifunctional polypeptides of vertebrate and fungal fatty acid synthases have arisen by independent gene fusion events. *FEBS Lett* **162**, 300-304.

Meier, P. & Wackernagel, W. (2005). Impact of mutS inactivation on foreign DNA acquisition by natural transformation in Pseudomonas stutzeri. *Journal of Bacteriology* **187**, 143-154.

Menendez, N., Nur-e-Alam, M., Brana, A. F., Rohr, J., Salas, J. A. & Mendez, C. (2004). Biosynthesis of the antitumor chromomycin A3 in Streptomyces griseus: analysis of the gene cluster and rational design of novel chromomycin analogs. *Chem Biol* **11**, 21-32.

Metsa-Ketela, M., Salo, V., Halo, L., Hautala, A., Hakala, J., Mantsala, P. & Ylihonko, K. (1999). An efficient approach for screening minimal PKS genes from Streptomyces. *FEMS Microbiol Lett* **180**, 1-6.

Mochizuki, S., Hiratsu, K., Suwa, M., Ishii, T., Sugino, F., Yamada, K. & Kinashi, H. (2003). The large linear plasmid pSLA2-L of Streptomyces rochei has an unusually condensed gene organization for secondary metabolism. *Mol Microbiol* **48**, 1501-1510.

Moreno-Hagelsieb, G., Trevino, V., Perez-Rueda, E., Smith, T. F. & Collado-Vides, J. (2001). Transcription unit conservation in the three domains of life: a perspective from Escherichia coli. *Trends Genet* **17**, 175-177.

Muller, T. A., Werlen, C., Spain, J. & Van Der Meer, J. R. (2003). Evolution of a chlorobenzene degradative pathway among bacteria in a contaminated groundwater mediated by a genomic island in Ralstonia. *Environ Microbiol* **5**, 163-173.

Mushegian, A. R. & Koonin, E. V. (1996). Gene order is not conserved in bacterial evolution. *Trends Genet* **12**, 289-290.

Nandi, S., Maurer, J. J., Hofacre, C. & Summers, A. O. (2004). Gram-positive bacteria are a major reservoir of Class 1 antibiotic resistance integrons in poultry litter. *Proc Natl Acad Sci U S A* **101**, 7118-7122.

Nemergut, D. R., Martin, A. P. & Schmidt, S. K. (2004). Integron diversity in heavy-metal-contaminated mine tailings and inferences about integron evolution. *Appl Environ Microbiol* **70**, 1160-1168.

Nesvera, J., Hochmannova, J. & Patek, M. (1998). An integron of class 1 is present on the plasmid pCG4 from gram-positive bacterium Corynebacterium glutamicum. *FEMS Microbiol Lett* **169**, 391-395.

O'Connor, S. (2004). Aureolic acids: similar antibiotics with different biosynthetic gene clusters. *Chem Biol* **11**, 8-10.

Opel, M. L., Arfin, S. M. & Hatfield, G. W. (2001). The effects of DNA supercoiling on the expression of operons of the ilv regulon of Escherichia coli suggest a physiological rationale for divergently transcribed operons. *Molecular Microbiology* **39**, 1109-1115.

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D. & Maltsev, N. (1999). The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 2896-2901.

Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Kyrpides, N., Fonstein, M., Maltsev, N. & Selkov, E. (2000). WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Research* **28**, 123-125.

Overbeek, R., Larsen, N., Walunas, T. & other authors (2003). The ERGO (TM) genome analysis and discovery system. *Nucleic Acids Research* **31**, 164-171.

Pal, C. & Hurst, L. D. (2004). Evidence against the selfish operon theory. *Trends Genet* **20**, 232-234.

Partridge, S. R., Brown, H. J. & Hall, R. M. (2002). Characterization and movement of the class 1 integron known as Tn2521 and Tn1405. *Antimicrob Agents Chemother* **46**, 1288-1294.

Piepersberg, W. (1993). Streptomycetes and corynebacteria. In *Biological fundamentals*, pp. 434-468. Edited by H. Sahm. Weinheim: Verlag Chemie.

Piepersberg, W. (**1997**). Molecular biology, biochemistry, and fermentation of aminoglycoside antibiotics. *Biotechnology of antibiotics*, 2nd ed. p. 81-163. Edited by W. R. Strohl. Marcel Dekker, Inc., New York, N.Y.

Piepersberg, W. (2002). Endogenous antimicrobial molecules: an ecological perspective. In *Molecular medical microbiology*, pp. 561-584. Edited by M. Sussman. San Diego: Academic Press.

Ploy, M. C., Courvalin, P. & Lambert, T. (1998). Characterization of In40 of Enterobacter aerogenes BM2688, a class 1 integron with two new gene cassettes, cmlA2 and qacF. *Antimicrob Agents Chemother* **42**, 2557-2563.
Price, M. N., Huang, K. H., Arkin, A. P. & Alm, E. J. (2005). Operon formation is driven by co-regulation and not by horizontal gene transfer. *Genome Res* **15**, 809-819.
Reams, A. B. & Neidle, E. L. (2004). Selection for gene clustering by tandem duplication. *Annu Rev Microbiol* **58**, 119-142.
Recchia, G. D. & Hall, R. M. (1995). Gene cassettes: a new class of mobile element. *Microbiology* **141 (Pt 12)**, 3015-3027.
Rhee, K. Y., Opel, M., Ito, E., Hung, S. P., Arfin, S. M. & Hatfield, G. W. (1999). Transcriptional coupling between the divergent promoters of a prototypic LysR-type regulatory system, the ilvYC operon of Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America* **96**, 14294-14299.
Rohr, J., Mendez, C. & Salas, J. A. (1999). The biosynthesis of aureolic acid group antibiotics. *Bioorganic Chemistry* **27**, 41-54.
Rosser, S. J. & Young, H. K. (1999). Identification and characterization of class 1 integrons in bacteria from an aquatic environment. *J Antimicrob Chemother* **44**, 11-18.
Schweizer, M., Roberts, L. M., Holtke, H. J., Takabayashi, K., Hollerer, E., Hoffmann, B., Muller, G., Kottig, H. & Schweizer, E. (1986). The pentafunctional FAS1 gene of yeast: its nucleotide sequence and order of the catalytic domains. *Mol Gen Genet* **203**, 479-486.
Shearwin, K. E., Callen, B. P. & Egan, J. B. (2005). Transcriptional interference - a crash course. *Trends in Genetics* **21**, 339-345.
Singleton, P. & Sainsbury, D. (2001). *Dictionary of microbiology and molecular biology*. Chichester: John Wiley and sons, Ltd.
Sneppen, K., Dodd, I. B., Shearwin, K. E., Palmer, A. C., Schubert, R. A., Callen, B. P. & Egan, J. B. (2005). A mathematical model for transcriptional interference by RNA polymerase traffic in Escherichia coli. *Journal of Molecular Biology* **346**, 399-409.
Springael, D. & Top, E. M. (2004). Horizontal gene transfer and microbial adaptation to xenobiotics: new types of mobile genetic elements and lessons from ecological studies. *Trends Microbiol* **12**, 53-58.
Stahl, F. W. & Murray, N. E. (1966). The evolution of gene clusters and genetic circularity in microorganisms. *Genetics* **53**, 569-576.
Stotzky, G. (1997). Soil as an environment for microbial life. In *Modern soil microbiology*, pp. 1-20. Edited by J. D. van Elsas, J. T. Trevors & E. Wellington. New York: Marcel Dekker, Inc.
Tiirola, M. A., Wang, H., Paulin, L. & Kulomaa, M. S. (2002). Evidence for natural horizontal transfer of the pcpB gene in the evolution of polychlorophenol-degrading sphingomonads. *Appl Environ Microbiol* **68**, 4495-4501.
Tolba (2004). Ph.D thesis. University of Warwick.
Tolba, S., Egan, S., Kallifidas, D. & Wellington, E. M. H. (2002). Distribution of streptomycin resistance and biosynthesis genes in streptomycetes recovered from different soil sites. *Fems Microbiology Ecology* **42**, 269-276.
Tringe, S. G., von Mering, C., Kobayashi, A. & other authors (2005). Comparative metagenomics of microbial communities. *Science* **308**, 554-557.
Vanden Boom, T. & Cronan, J. E., Jr. (1989). Genetics and regulation of bacterial lipid metabolism. *Annu Rev Microbiol* **43**, 317-343.

Vining, L. C. (1992). Roles of secondary metabolites from microbes. In *Secondary metabolites: their function and evolution*, pp. 184-198. Edited by D. J. Chadwick & J. Whelan. Chichester: John Wiley and sons.
Weber, T., Welzel, K., Pelzer, S., Vente, A. & Wohlleben, W. (2003). Exploiting the genetic potential of polyketide producing streptomycetes. *J Biotechnol* **106**, 221-232.
Wiener, P. (1996). Experimental studies on the ecological role of antibiotic production in bacteria. *Evolutionary Ecology* **10**, 405-421.
Wiener, P. (2000). Antibiotic production in a spatially structured environment. *Ecology Letters* **3**, 122-130.
Witkowski, A., Rangan, V. S., Randhawa, Z. I., Amy, C. M. & Smith, S. (1991). Structural organization of the multifunctional animal fatty-acid synthase. *Eur J Biochem* **198**, 571-579.

FIGURE LEGENDS AND FIGURES (BELOW)

Figure 1**. Biosynthetic gene clusters for streptomycin and bluensomycin:** streptomycin gene cluster from *Streptomyces griseus,* dihydroxystreptomycin cluster from *S. glaucescens,* partial cluster of bluensomycin from *S. bluensis,* bluensomycin gene cluster from S. bluensis as reported by Jung *et al*. (2003). The clusters are aligned according to their homolgous *strB1*. Adapted from Tolba (2004), after Piepersberg (1997) and (Jung *et al.*, 2003).

Figure 2. **Linear chromosome of Streptomyces**. Arm and core regions seem to contain genes for different metabolic functions. The linear chromosome structure might occasionally recombine from only one crossover after partial diploid formation to exchange a telomere protein-bound end with another chromosome fragment. However, the telomere protein ends might be bound together holding the chromosome in circular form. Adapted from Piepersberg (2002).

Figure 3. **Evolutionary pressures on streptomycin producers during anthropogenic streptomycin application**. All three cases (producers with full biosynthetic cluster, resistant organisms with some biosynthesis genes conferring resistance, and ribosome mutants) might be expected in isolates if, as Firn & Jones (2000, 2003) suggest, natural selection favours organisms able to produce and innovate secondary metabolites at little cost.

Fig 1

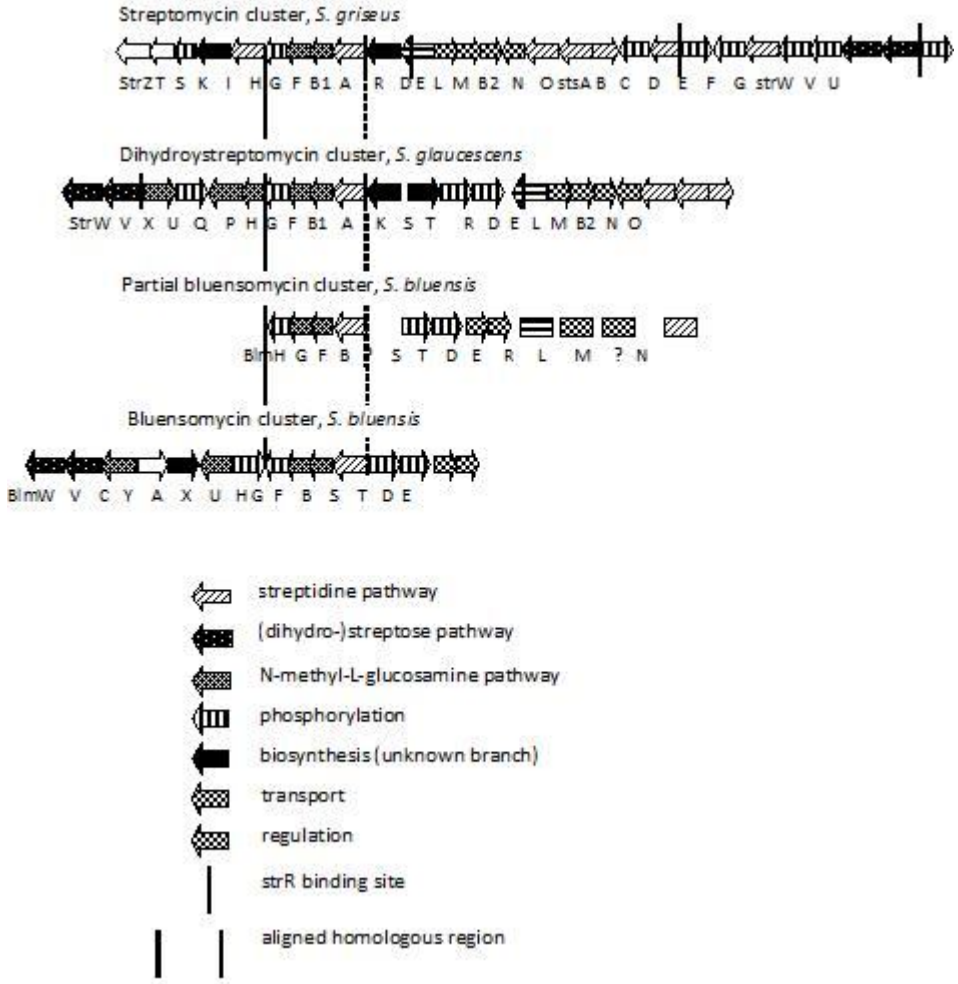

Streptomycin cluster, S. griseus
StrZT S K I H G F B1 A R D E L M B2 N O stsA B C D E F G strW V U
Dihydroystreptomycin cluster, S. glaucescens
StrW V X U Q P H G F B1 A K S T R D E L M B2 N O
Partial bluensomycin cluster, S. bluensis
BlmH G F B S T D E R L M ? N
Bluensomycin cluster, S. bluensis
BlmW V C Y A X U H G F B S T D E
streptidine pathway
(dihydro-)streptose pathway
N-methyl-L-glucosamine pathway
phosphorylation
biosynthesis (unknown branch)
transport
regulation
strR binding site
aligned homologous region

Fig 2

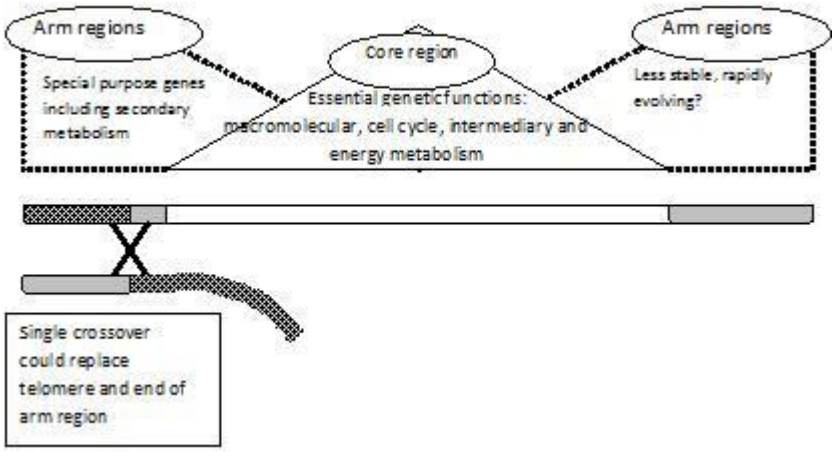
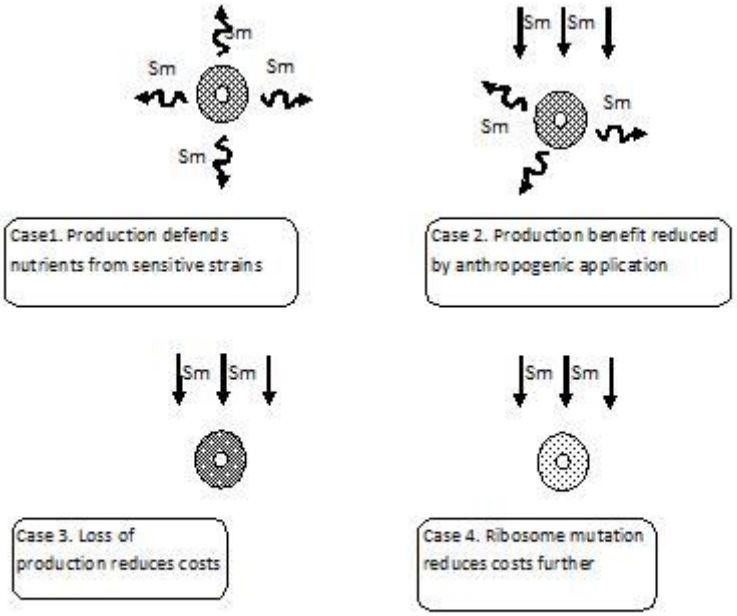

Arm regions
Special purpose genes including secondary metabolism
Core region
Essential genetic functions: macromolecular, cell cycle, intermediary and energy metabolism
Arm regions
Less stable, rapidly evolving?
Single crossover could replace telomere and end of arm region

Fig 3


Sm
Sm
Sm
Sm
Sm
Sm
Sm
Sm
Sm
Sm
Case1. Production defends nutrients from sensitive strains
Case 2. Production benefit reduced by anthropogenic application
Case 3. Loss of production reduces costs
Case 4. Ribosome mutation reduces costs further